THE UNIVERSITY
of EDINBURGH

COMPUTATIONAL METHODS FOR THE ANALYSIS OF NON-CELL-AUTONOMOUS PHENOMENA AND DERIVED GENE CO-EXPRESSION NETWORKS

SAMUEL HERON

Doctor of Philosophy
School of Informatics
University of Edinburgh

2018

Samuel Heron:
Computational Methods for the Analysis of Non-Cell-Autonomous Phenomena and Derived Gene Co-expression Networks
Doctor of Philosophy, 2018

SUPERVISORS:
T. Ian Simpson
Giles Hardingham
Owen Dando

LAY SUMMARY

'Non-cell-autonomous effects' (NCAE) are changes observed in one cell or cell-type as a consequence of the actions of another. The study of these effects is crucial to our understanding of how cells behave when they exist together as part of a tissue (e.g. white matter tissue in the brain). Investigation of NCAE have been greatly aided by advances in sequencing technologies that provide us with information about which genes in a cell are active at the time of sequencing. This information can be used to investigate changes to cell behaviour, such as those brought about by NCAE.

To study these effects on cells we have to get a separate set of genetic information from each cell-type we are investigating in order to be certain that changes we see are a result of interaction with other cells. Previously this has been done by physically separating the cells, however this causes the cells a great deal of stress and many to die. This is a problem as it means that the genetic information extracted from these cultures is then biased towards genes that activate in response to stress or cell-death. Such physical methods also risk bias by incompletely separating the cells.

In this thesis I introduce a new computational tool that assigns sequence information by cell-type without the need to physically separate the cells. This is done by using cultures where each cell type belongs to a different closely related species, for example one cell-type being from mouse and another from rat. Sequenced information is then compared to the genomes of each species and assigned to the one it matches best. As a result we get two separate datasets, one belonging to each species, that each contain the information for a separate cell-type allowing us to study the genetic changes resulting from the interaction of these cell-types: NCAE.

Analysis of this data is frequently conducted through the use of network models, where a node represents a gene and a connecting edge represents how the strength of one gene's activation correlates with another over several experimental conditions. These networks allow us to examine changes in genetic information across experimental conditions to identify what biological processes are occurring, for example those as a consequence of NCAE. This is frequently carried out by clustering the network and examining the clustered genes. In this thesis I introduce a new tool, that uses 'entropy' (a statistical measure of disorder) to examine the strength, or order, in the representation of genes belonging to specific biological processes.

Lastly I apply these new tools to two mixed-species datasets, examining the effect of different stimuli on NCAE between different cell-types in the brain.

ABSTRACT

Non-cell-autonomous effects are the changes observed in one cell or cell-type as a consequence of the actions of another. The study of these phenomena is crucial to our understanding of how diverse cell-types function and co-operate together in complex tissues. The investigation of these effects has been greatly advanced by the advent of next-generation sequencing (NGS) technologies which enable the rapid sequencing of genetic information. NGS data, such as RNA-Seq, can be analysed computationally to allow comparison of cellular transcriptomes. In practice, the study of non-cell-autonomous phenomena through NGS has relied upon the physical separation of cell populations in order to be sure that derived transcriptomic data is exclusively from one cell type or the other. However these methods have been shown to introduce noise as a result of stress induced by the separation process, whilst also being susceptible to bias through contamination resulting from imperfect separation of cell populations. In this thesis, a pipeline was developed to provide an *in silico* means of investigating these phenomena without the need for physical separation. The pipeline takes RNA-seq reads from novel mixed-species populations - *in vitro* cultures where each cell type is derived from a distinct species - and sorts them according to species specific origin using quality variables from multiple genome mappings as discriminators. Our method is demonstrably robust to incorrect assignment and shows high precision and recall across species of differing genetic distances, thereby providing an alternative to flawed physical separation techniques. Downstream study of such RNA-seq samples is increasingly conducted using network methodologies. Gene co-expression networks have been demonstrated as a biologically representative means for analysing NGS data. However, many existing methods for attributing the involvement of biological function to networked datasets disregard the structural information provided within them. In this thesis, I build upon an existing approach to use information theoretic entropy as a method for network-based enrichment and thereby demonstrate that the integration of network edge information can be used to more reliably infer biological pathway involvement. Our method out-performs the original whilst correcting for pathway-size bias. Lastly, the utility of the methods presented in this thesis was demonstrated through application to the study of two different phenomena: the induction of neural activity on co-cultures of neurons with astrocytes and the stimulation of microglia by LPS on co-cultures of microglia, neurons and astrocytes, by investigating cell-type specific involvement of biological pathways.

ACKNOWLEDGEMENTS

I would like to thank my supervisors T. Ian Simpson, Giles Hardingham and Owen Dando for the their tireless support and advice throughout my PhD. My weekly meetings with Ian and Owen were an amazing opportunity to talk through and exchange ideas, always serving to keep me on track and to never struggle under the weight of my problems, without which this work would not be what it is. Ian has imparted not just his knowledge and help as a supervisor but has been an inspiring role model as a researcher, one whose enthusiasm and work ethic I hope to emulate in myself. Working with Owen on the Sargasso project has been an experience of great insight as a programmer and has demonstrated to me the effectiveness of collaborative coding. Both Giles and Owen have provided indispensable help and support with relevant neuroscientific knowledge throughout the project. I would also like to extend my thanks to Philip Hasel and Jing Qui whose work with Giles has created the data that has made my work possible. Lastly I cannot thank Ian and Owen enough for their repeated proof readings of my thesis during these final months. I could not have asked for more from all my supervisors, their advice, encouragement and support throughout my PhD has been invaluable and for it I am truly grateful.

I would also like to thank the staff and cohorts of the Neuroinformatics and Computational Neuroscience DTC, the former who have trained me in this field and provided me with the knowledge to embark on my research and the latter who made this process so enjoyable, for the camaraderie and the mutual support we enjoyed. I would like to thank those in my DTC13 cohort for their help and support, in particular Alina Selega and Gavin Gray for their continued friendship and support. In addition to this the members of my research group have been a joy to work and spend time with, indeed our long board gaming nights have provided a great social compliment to our work.

Lastly outside of my academic circles, the support from my friends who have been through the PhD process has gone a long way to helping me find my way through it and our conversations have done much for my confidence, I would like to thank Alina, Gavin, Maciej Pajak, Emma Brown Dewhurst, Joe Dewhurst and Hamish Kallin in particular. The support of my parents and relatives throughout this time has also been a great boon to my well-being. I would finally like to express my tremendous gratitude to my partner Mari Lehvä, without whose emotional support I would not have got this far and who has been there for me every step of the way.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, 2019

Samuel Heron, June 28, 2019

CONTENTS

ABBREVIATIONS

3SPcc	Three Species Co-culture (Dataset, see Section 1.4.3)
3SPcc-HA	Three Species Co-culture Dataset, separated Human Astrocyte data
3SPcc-MN	Three Species Co-culture Dataset, separated Mouse Neuron data
3SPcc-RM	Three Species Co-culture Dataset, separated Rat Microglia data
AD	Activity Dependence (Dataset, see Section 1.4.2)
ADcc-MA	Activity Dependence Co-culture, separated Mouse Astrocyte data
ADcc-RN	Activity Dependence Co-culture, separated Rat Neuron data
ADm-MN	Activity Dependence Monoculture (Mouse Neuron)
ADm-RN	Activity Dependence Monoculture (Rat Neuron)
AEN	All gene Expression Network
AS	Alignment Score
bp	Base Pairs, for sequence length
BH	Benjamini-Hochberg procedure
BiC	Bicuculline (Receptor agonist)
cDNA	complimentary DNA
DE	Differential Expression
DNA	Deoxyribonucleic acid
DNeA	Differential Network Analysis
DiNA	Differential Network Analysis (Tool) (Gambardella et al., 2013)
DW	Differential Wiring
FACS	Fluorescence Activated Cell Sorting
FISH	Fluorescence in situ Hybridization
FPKM	Fragments Per Kilobase Million

HMM	Hidden Markov Model
LCM	Laser Capture Microdissection
LPS	Lipopolysaccharide (Inflammatory molecule derived from gram-negative bacteria)
LTP	Long-term Potentiation
mRNA	Messenger RNA
MTC	Multiple Testing Correction
NCAE	Non-cell-autonomous Effects
NGS	Next Generation Sequencing
OS	Oxidative Stress (Dataset, see Section 1.4.1)
OSm-MA	Oxidative Stress Monoculture (Mouse Astrocyte)
PAN	Immunopanning
PCN	Protein Coding gene Network
PCR	Polymerase Chain Reaction
PE-A	Pathway Entropy 'all' Approach
PE-Aw	Pathway Entropy 'all' Approach, using weighted information
PE-Au	Pathway Entropy 'all' Approach, unweighted (using edge counts)
PE-T	Pathway Entropy 'topology' Approach
PE-Tw	Pathway Entropy 'topology' Approach, using weighted information
PE-Tu	Pathway Entropy 'topology' Approach, unweighted (using edge counts)
rRNA	Ribosomal Ribonucleic acid
RNA	Ribonucleic acid
RPKM	Reads Per Kilobase Million
scRNA-Seq	Single Cell RNA Sequencing
TBOA	DL-threo-b-Benzyloxyaspartic acid (glutamate transporter inhibitor)
TRAP	Translating Ribosome Affinity Purification

TTX	Tetrodoxin (voltage-gated Na^+ channel blocker)
WGCNA	Weighted Gene Co-expression Network Analysis (Tool) (Langfelder and Horvath, 2008)

1

INTRODUCTION

1.1 MOTIVATION

All lifeforms, exceeding a sufficient level of complexity, are comprised of cellular tissues whose functions requires the cooperation of multiple specialised cellular types. Over the last century, endeavours in the life sciences have isolated many key cells for study, characterising their behaviour and function mostly in isolation. However these properties form an incomplete picture of the roles played by many cell-types, particularly those as complex as are found in the mammalian brain, such as neurons. These cell-types, when grown isolated from other cell-types through *in vitro* monoculture, do not develop many of the attributes we observe *in vivo*. Whilst some of this difference may be simply a result of the non-cellular aspects such different environments (e.g chemical composition), study has shown that when neurons are co-cultured with astrocyte glia, cells which closely cooperate with neurons through metabolic support, neural characteristics better resemble those of *in vivo* cells. It is these interactions that take place between cells, so called 'non-cell-autonomous effects' (NCAE), that we are only now just beginning to fully explore with the sufficient advance of technology.

A more comprehensive understanding of NCAE would fundamentally increase our core knowledge of cellular function and hold great potential for medical and particularly drug research through allowing better modelling of downstream interactions between cells and within tissues. It also holds potential for improving *in vitro* co-culture of cells whose accuracy, in terms of approximation of cell function *in vivo*, is often far from perfect.

Investigations conducted into NCAE on gene expression usually require the co-culture of cells of interest *in vitro* as extraction of a limited number of cell-types from *in vivo* subjects risks biasing results through the uncontrolled presence of other cell-types. These cultures are then physically separated and sequenced for computational analysis to determine changes in gene expression and activity of biological pathways. However, physical separation has been demonstrated to add noise and bias the data obtained (Okaty et al., 2011b) rendering a key means by which NCAE are studied as imperfect.

The research in this thesis is motivated by advancements in genomics and computational methods which we believe can be better utilised to improve upon the existing methods for studying NCAE and indeed circumvent present problems.

1.2 BIOLOGICAL BACKGROUND

There are several key areas of biological research integral to the work presented on NCAE in this thesis. I shall firstly describe the sequencing of RNA, through which the core datasets for our investigation are obtained. I will then proceed to describe the culturing of cells *in vitro* for the purposes of NCAE, the effect this has on development and maturation compared to *in vivo* and the benefits, in addition to enabling NCAE study, of co-culturing cell-types. Lastly I will summarise the current means by which datasets containing multiple cell-types are separated for NCAE investigation.

1.2.1 *Next Generation Sequencing & RNA-Seq*

Next generation sequencing (NGS) (Schuster, 2007) has greatly improved our ability to understand and analyse the genome and the sources of transcriptomic change. Through increasing speed and accuracy whilst decreasing the cost of its predecessor, Sanger sequencing, NGS approaches have vastly increased feasibility and diversity of sequence data production in computational biology.

Various different NGS technologies exist, however for Illumina sequencing, which has been used to generate the sequence data used in this thesis, a typical procedure is as follows: the sample input for sequencing, for example complementary DNA (cDNA), is cleaved into small sections which are amplified through the application of the polymerase chain reaction (PCR), these amplified sequences are then separated into individual strands, the resulting millions of sequences are then attached to the flow cell so that they can be sequenced at the same time. The sequencing process involves flooding the separated strands with *DNA polymerase* enzymes and free nucleotides, colour coded by base using fluorescent tags, which have a terminator preventing the addition of more than one nucleotide by DNA polymerase. An image is then taken and this process is then repeated adding a single nucleotide each time. Subsequent computational analysis then determines the sequences through identification of the fluorescent tag for each read in each image (EMBL-EBI, 2017). This process is illustrated in Figure 1.1.

RNA-Seq is an approach for the specific sequencing of RNA through the use of NGS technology. The use of RNA-Seq entails the reverse transcription of RNA present in a biological sample, through the use of the enzyme *reverse transcriptase*, to generate

a library of cDNA. This cDNA library is then sequenced through the application of an NGS approach (Wang et al., 2009), as described above. The result of this sequencing is a dataset of RNA-Seq 'reads', typically between 50 and 150 base pairs (bp) in length (dependent on NGS technology) and paired-end, that can then be aligned to the reference genome of the appropriate species for further study. Sequencing depth for RNA-Seq studies is usually measured per million reads sequenced, with a high depth being important to ensure full representation of low-expressed genes. Whilst this process is still the most commonly used, the conversion of RNA to cDNA is a potential locus for error and bias (Marguerat and Bähler, 2010). In response approaches to directly sequence the RNA molecules without conversion have arisen (Ozsolak et al., 2009).

To facilitate NGS sequencing, the RNA or resultant cDNA libraries must undergo fragmentation to reduce their size (Wang et al., 2009). This introduces a bias towards gene length when it comes to analysing the abundance of expression as larger genes will generate higher read counts thus prohibiting the direct comparison of counts. To combat this normalisation steps are to be taken when dealing with the raw RNA read counts such as the calculation of reads per kilobase million (RPKM) (Mortazavi et al., 2008) or fragments per kilobase million (FPKM) (Trapnell et al., 2010). This normalisation step involves dividing the total reads in a sample by 10^6 (to normalise for sequencing depth) and scaling gene counts down by this factor; gene counts are then divided by gene length (to normalise for gene length). FPKM has been used when looking at abundance in this thesis as our data is paired-end RNA-Seq where each read is sequenced from either end, which FPKM takes into account by not counting reads derived from the same fragment individually, but as pairs.

RNA-Seq thus captures a snapshot of the presence and quantity of RNA for a given biological sample at the point of sequencing, therefore allowing us to examine the expression of genes at the point of sequencing. Its use in comparing pre and post-stimulus samples can greatly inform us of a given stimulus' effect on the regulation of genes in a particular experimental context. In this thesis we are primarily interested in investigating the changes in gene expression observed in one cell-type as a result of interaction with a different cell-type.

Whilst analysis of gene expression change has often been conducted through the use of differential expression analysis (Love et al., 2014), where the change in expression of a gene is examined across experimental conditions in order to deduce whether expression undergoes significant change between conditions, more frequently now more complex computational methods, such as differential network analysis (Gambardella et al., 2013), are being employed to identify such changes across groups of co-expressed genes.

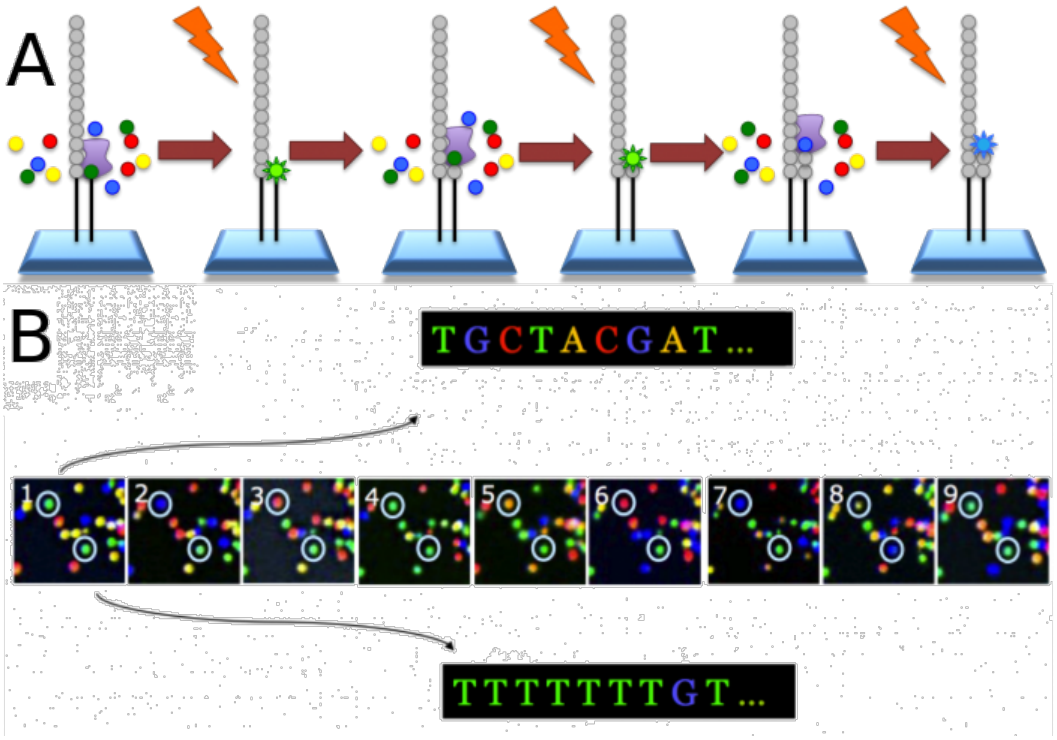


Figure 1.1: **Illumina Sequencing Technology**. This core steps in Illumina's sequencing process: (A) Reads attached to a slide are flooded with fluorescently labelled nucleotides with terminators; these are attached one at a time by DNA polymerase and an image is taken. This step is then repeated for each nucleotide. (B) The images taken at each binding step are then analysed by computer imaging to determine the nucleotide attached in each image and the resultant read sequences are output to file. Figure reproduced from (EMBL-EBI, 2017).

1.2.2 *Neural Cell Types Used in This Study*

The work in this thesis uses genetic data derived from several key neural cells: neurons, astrocytes and microglia. In this section I will provide a brief description of each of these cell types and an overview of their known functions in brain.

1.2.2.1 *Neurons*

Neurons are a highly specialised cell-type found in the brain and the central nervous system. Physiologically they are comprised principally of a cell body, an axon and dendrites. Dendrites are branching structures that protrude from the cell body, these tree-like structures receive input from synaptic junctions with other neurons which, if sufficiently strong, enables a neuron to transmit an electrical signal down its axon. Such a signal will in turn stimulate other neurons via synapses positioned after the axon. Physiology in neurons can vary depending on sub-type according to functional role.

Signaling occurs at cell-junctions between neurons known as synapses, these junctions make use either of electrical gradients or of chemical components to send signals. Synapses can strengthen or weaken depending on use and thus the activation of a network of neurons through specific synapses can connote a response to a particular stimuli; for example recognition of a specific scent in the brain's olfactory bulb.

Due to the highly plastic and programmable nature of synapses, neurons can thus play a range of different roles; from interpreting sensory stimuli to the storage or memories, and thus have been a primary focus of neuroscientific research since the establishment of the field.

1.2.2.2 *Astrocytes*

The brain and central nervous system are principally comprised of neurons and glial cells of which astrocytes are a sub-type of the latter. Astrocytes derive their name from the star-like shape of the cellular processes that protrude from the central cell body and are the most common form of glial cell within the brain. As they work closely with neurons, the cellular processes of astrocytes are observed to envelop synaptic junctions between neurons, this proximity aiding their supportive functions.

The proportion of astrocytes is variable in the brain from region to region and their role is typically understood as one of support, particularly metabolic support and through repair. However they have been observed to play many additional roles, from aiding the regulation of blood flow to modulating synaptic transmission between neurons, which has made them a subject of more frequent and intense study.

1.2.2.3 *Microglia*

Microglia are another form of glial cell found throughout the brain and the central nervous system. Physiologically microglia are plastic and can change their form to suit function as required, for example to become macrophages in response to infection.

Microglia play several key roles in the brain, perhaps most importantly is their immunological role in response to infection which involves promotion of inflammation in affected areas and inter-cellular signaling, particularly through the use of cytokines. They have also been observed to function similar to macrophages through use of phagocytosis and presentation of antigens. In addition to this microglia can promote repair, carry out maintenance in the area surrounding them by removing dead cells and plaques as well as through the ability to remove synapses between neurons.

1.2.3 *Investigation of Non-Cell Autonomous Phenomena*

Non-cell-autonomous effects (NCAE), simply, are changes brought about in one cell as a result of the action of another distinct cell. Research (Johnson et al., 2007; De Luca et al., 2015; Ivanov et al., 2015) is often directed at investigating the interactions of different cell-types in order to better characterise our understanding of the role and behaviours of specific cell-types, knowledge important for areas from understanding disease and drug development to gene regulation and optimisation of *in vitro* culturing. This area of study often therefore takes place *in vitro* where we can have greater control over variables relating to cell development and maturation and can more easily restrict the context of study to only those cells we are interested in. Indeed, it is *in vitro* data that I work with throughout this thesis. To analyse the data produced by these cultures, we must find a method by which we can separate the data from each cell-type so that we can confirm whether gene expression changes are indeed the result of NCAE. This is usually carried out through the application of physical separation techniques.

In this section I will therefore first describe the difficulties of culturing of cells *in vitro* before moving on to specifically detailing research regarding the co-culturing of distinct cell-types. This will be done with a focus on neural cell-cultures as that is the domain of biological focus in this thesis. Lastly I will overview the separation methods currently in use, both physical and *in silico*.

1.2.3.1 *Co-Culture of Distinct Cellular Types*

Stem cell co-culturing, the intentional directed culturing of two or more distinct cell types, has become both increasingly viable, with the improvements in directed differentiation mentioned in the previous section, and increasingly attractive for the better *in vitro* culturing of cells and the study of non-cell-autonomous effects. Whilst this kind of culturing has happened unintentionally for almost as long as we've been culturing stem cells, due to our lack of control of stem cell differentiation, it has more recently become useful particularly in neural cultures as a method of better recreating *in vivo*-like environmental conditions, for example the inclusion of glial cells to support neurons. Glial cells have been previously established as a support cell for neurons, providing structural and metabolic support as well as encouraging synaptogenesis (Christopherson et al., 2005) and thus in this context have been shown to improve development and maturation in neurons when co-cultured together *in vitro* (Johnson et al., 2007). Of particular interest, such co-culture methodology has demonstrated a decrease in undifferentiated stem cells (Johnson et al., 2007). Both of these effects thus highlighting the importance of NCAE.

Experimental studies have employed co-culture for a variety of different purposes, several of which will be described here in order to highlight the versitality of study using *in vitro* co-culture to investigate NCAE. A methodology for three dimensional *in vitro* co-culturing has been used to assess the impact of cell invasion or disease on a tissue of cells (Chintala et al., 1997; Nygaard et al., 1994; Amann et al., 2014). A recent study (Ivanov et al., 2015) used this methodology to culture human neurons and medulloblastoma cells, a common type of brain tumour, from stem cells into 3-D 'spheroids', roughly spherical conglomerations of the cultured cells. Cells were then stained with a combination of different dyes in order to assess cell type and to mark dead cells. The effect of the medulloblastoma tumour on the health and organisation of the cultured neurons could then be assessed using 2-photon microscopy. Such co-culturing has been used as an attempt to recreate a known or observed *in vivo* biological phenomena *in vitro* in order to better understand the actions of the cells involved (Daverey et al., 2014; Li and Wang, 2013). Indeed (De Luca et al., 2015) did this by co-culturing dorsal root ganglia (DRG) neurons and a Schwann cell phenotype, derived from adipose-derived stem cells, in order to study myelination and nerve regeneration in the peripheral nervous system. This study aimed to recreate the interaction and communication of these cells *in vitro* in the hope that this experimental methodology could be used as a model for understanding the known *in vivo* interaction.

With the continued study of NCAE in this manner it is hoped that we are drawing closer to an era of clinically viable *in vitro* tissue experimentation as our ability to accurately engineer complex cell cultures improves in both accuracy and diversity. The idea of transferring molecular biology study, particularly human disease study, from *in vivo* animal models to *in vitro* human cultures is a very real prospect that may result from continued work into NCAE.

1.2.3.2 *Physical Separation Techniques*

When studying NCAE *in vitro* it is necessary to separate the cell types of study otherwise we cannot assign the derived gene expression to its cell-type of origin. This is presently done primarily through the use of physical separation techniques such as laser capture microdissection (LCM), translating ribosome affinity purification (TRAP), fluorescence activated cell sorting (FACS), immunopanning (PAN), and manual sorting.

LCM involves the use of a laser to cut around and separate an area of interest in a culture (Espina et al., 2006) as can be seen in Figure 1.2. TRAP utilises cells modified to express a cell-type specific transgene, such as one producing a fluorescent protein, in its ribosomes in order to indirectly target mRNA. When mRNA is collected at the

point of sequencing, any that has tagged ribosomes attached can be separated in a cell-type specific manner; this means only mRNA that was in the process of translation is sequenced (Heiman et al., 2014). FACS involves the sorting of cells based upon the fluorescent characteristics of a cell (Herzenberg et al., 2002); this can be determined by the expression of a particular fluorescent protein for example. PAN involves the use of specific antibodies to physically bind to and thus separate a cell-type of interest (Cahoy et al., 2008). For manual sorting, cells are fluorescently tagged and manually collected by aspiration through the use of fluorescence stereomicroscope and pipette (Hempel et al., 2007).

Whilst all these methods do separate the cultures of study to enable investigation of NCAE, it has also been demonstrated that all methods risk contamination, though particularly LCM and TRAP, in addition to PAN introducing bias to gene expression through over-activation of stress and cell-death responses (Okaty et al., 2011b,a). The cellular damage caused by LCM can be seen in Figure 1.2B. Whilst FACS, which is widely used for this kind of separation, exerts less physical pressure on cells, research has demonstrated that its use induces the expression of multiple immediate early genes; the use of PAN and simple trypsinization also trigger a similar effect (Hasel et al., 2017). This expression bias in addition to the risk of contamination, even if lower for FACS than for other methods (Okaty et al., 2011a), is concerning and will affect any downstream analysis that makes use of derived expression information, such as differential expression analysis. However, whilst this is known as a necessary evil of using these methods, a physical separation technique free of these risks is yet to be devised.

Spatial Transcriptomics & Single Cell RNA-Seq

Physical approaches such as spatial transcriptomics, particularly those based on single-cell RNA-Seq (scRNA), present an alternative means by which expression information can be physically separated in a tissue.

Spatial transcriptomics refers to a group of approaches by which expression can be measured at regular physical intervals across a single tissues.There are broadly two main approaches, those that make use of fluorescence in situ hybridization (FISH) and those based on scRNA (Burgess, 2019). In FISH-based methods, fluorescent markers are used to highlight transcript expression within the tissue. This method is limited both by the number of available fluorescent markers and the spatial overlapping of fluorescent signal when a large number of transcripts are visualised. As a result the transcriptomic resolution of FISH methods is more limited than RNA-Seq based methods. Advances are being made to increase this resolution however and perhaps

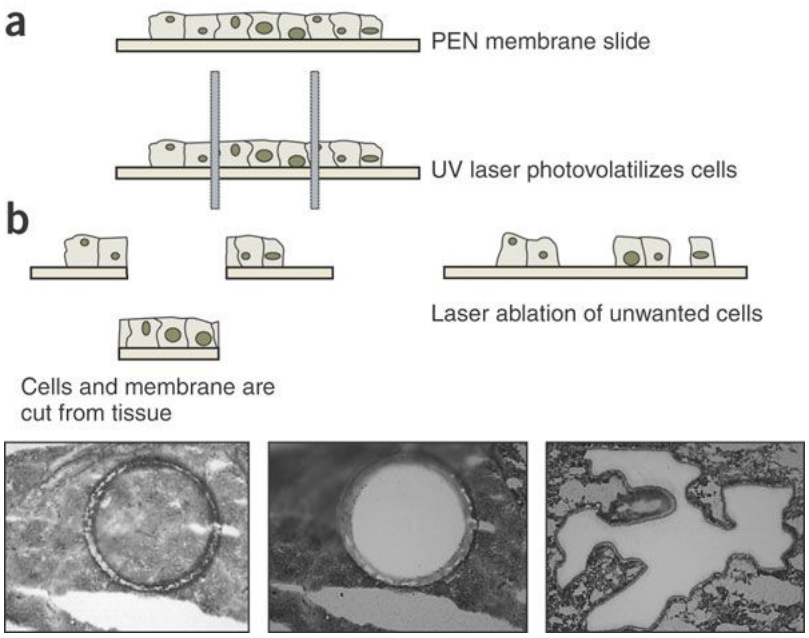


Figure 1.2: **Laser Capture Microdissection Procedure.** LCM carried out on a cell sample: (A) Tissue is mounted on a polyethylene napthalate (PEN) or polyethylene tetraphthalate (PET) membrane. A UV laser can be used to cut away cells of interest or to ablate unwanted tissue, leaving cells of interest intact on the substratum. (B) Veritas UV cutting tools allow defined circular cutting areas or freeform polygon areas to be microdissected. Figure and caption reproduced from (Espina et al., 2006).

in the future, this approach could be applied to thin tissue slices for a wide study of NCAE.

The use of scRNA approaches for spatial transcriptomics, in contrast, has no problem with transcriptomic resolution. For these approaches however, cells are dissociated from the tissue in order to be sequenced which hampers precise spatial reconstruction as information as to their exact original location is lost. A recent approach (Rodriques et al., 2019) solves this problem by taking tissue samples at an almost cellular resolution and digesting these for sequencing. Whilst this tackles both issues of spatial and transcriptomic resolution, that it takes 'tissue' samples means that the resulting RNA-Seq cannot be guaranteed to be from a specific cell-type, a precision necessary for the study of NCAE between cell-types.

1.2.3.3 *in silico Separation Techniques*

Whilst there are *in silico* techniques that can be applied to study NCAE, they are primarily specific to certain areas of research, such as metagenomics and xenograft studies. Given that they have been designed to work with data specifically from these areas, application outside of this to general purpose mixed cell-type data is limited.

I will however provide a brief overview of the approach these methods use. Lastly I will discuss the application of deconvolution based approaches to transcriptomics.

Separation Methods in Metagenomics

Metagenomics is a field that aims to identify species present in a multi-species dataset and partition the sequence information within it between those species. This taxonomic classification is known within the field as 'binning'. The application of these methods is regularly to samples taken from microbiomes, particularly the gut (Qin et al., 2012) and vaginal (Aagaard et al., 2012) microbiome in humans, in order to identify the microbial species present. However its application is quite broad and as such it has been used to identify species in aqueous (Xie et al., 2011) and soil (Fierer et al., 2012) environments. As the data for this thesis will require the partitioning of mixed-species data via genome mapping this field is a key area of comparison.

As metagenomics is a relatively young discipline within bioinformatics there is as yet no consensus on the best method to use for species identification and data separation, however two approaches are most prominently used: a method based on the use of phylogenetic markers or one based upon shotgun sequencing. I will briefly summarise these approaches before highlighting their recent applications and the challenges they face.

The phylogenetic classification method focuses on the extraction and sequencing of amplicons, the product of PCR-type experimental methods, that represent key phylogenetic markers depending on the kingdom of species targeted. These markers are sufficiently well conserved such that they can be used for taxonomic classification. For example if investigating fungal species amplicons representing genomic regions such as the 'internal transcribed spacer region' would be targeted (Santamaria et al., 2012), else if we were looking at prokaryotes we would be instead best looking for well conserved phylogenetic marker genes: rpoB or 16S rRNA for example (Clarridge, 2004). These markers are particularly useful as phenotypic analysis can be confounded due to inter-species gene sharing.

The shotgun sequencing method (Tyson et al., 2004) instead involves the extraction and sequencing of genomic fragments from the sample into reads. The resulting dataset comprises of many short reads that can be further analysed, for example by examining GC content and read depth, to uncover the properties of the genomes present in the original sample. This data is then run through an assembler to assemble the reads into scaffolds, informed and validated by any 16S rRNA genes - highly conserved genes in microbiota - in the data. These scaffolds can then be analysed in terms of length and conservation with known species in order to classify the data

or identify new species. This method, as with the previous, is however very dependent on the presence of 16S rRNA within shotgun samples in order to successfully assemble and identify species present and as such will potentially run into difficulty when this is not present in high enough quantities. Previously known information relating to the microbial communities present in the metagenomic dataset is also a significant contributor to its success regarding knowledge particularly of other phylogenetic markers.

Separation Methods in Xenograft Studies

Xenograft studies involve the transplantation, or 'grafting', of the cells or tissue from one species onto the cells or tissue of another. As a method it is particularly useful for the study of disease response and for cancer research as diseased or tumorous cells from human can be grafted into a healthy mouse model in order to analyse response a disease progression. In addition to the transplantation of human cells into animal models for study, xenografting of animal organs into human recipients is also an area of study.

When investigating gene expression in these samples however, there is the issue of cross species contamination of genetic data. Whilst physical separation techniques, as described in Section 1.2.3.2 have been used to tackle this issue, recently *in silico* approaches, such as 'Xenome' (Conway et al., 2012) and 'Disambiguate' (Ahdesmäki et al., 2016), have emerged as an alternative.

The Xenome approach constructs reference sets of k-mers from each species' reference genome, that of the host species and the graft species, and uses a set classification approach to determine whether individual RNA-Seq reads belong to either species, both, neither or are ambiguous.

Disambiguate's methodology on the other hand is to map reads to the genomes of both species, using the `TopHat` or `STAR` read aligner, and then to assign each read to either species using a pre-defined strategy, either: for `STAR`, the overall alignment score as primary discriminator and edit distance to break any ties, or for `TopHat`, the sum of the edit distance, reported alignments and quantity of gap opens is used (Ahdesmäki et al., 2016).

Despite the tool's difference in methodologies, their comparison in (Ahdesmäki et al., 2016) shows quite similar performance for the tested data. Whilst these tools could in theory be applied to more general mixed-species separation, their design and subsequent applications focuses understandably on the context of their intended use: two species host and graft scenarios.

Use of Deconvolution

Deconvolution methodologies broadly attempt to attribute information in heterogenous datasets to singular sources of origin. When applied to transcriptomics this would mean attributing expression information from a dataset of mixed cellular expression to their cell-types of origin. Various mathematical methodologies have been applied by deconvolution approaches, with the most common approaches utilising: least squares, support vector regression, dimensionality reduction techniques or non-negative matrix factorisation (Avila Cobos et al., 2018).

Deconvolution methods applied to transcriptomics frequently require prior knowledge of the expression profiles, or marker genes, of the cells present in the data. Whilst this means that they can be very effective at estimating cell-type specific abundances, for example CIBERSORT's deconvolution of leukocyte expression (Gentles et al., 2015), their reliance on prior knowledge hinders detection of *de novo* NCAE, particularly if such signals are not strongly expressed in the data. Similarly these methods are vulnerable to presence of unanticipated cell-types which will present a source of bias. Some of these methods also assume cellular expression profiles to be uncorrelated and do not account for effects introduced by different experimental stimuli, techonologies or platforms (Avila Cobos et al., 2018).

Thus whilst these methods can be useful for correlating expression with abundancies and measuring changes in cellular abundancies between samples or conditions, the fact that they do not produce a full resolution of cell-type specific expression is not ideal for the study of NCAE.

1.2.4 *Overview of Stem Cell Biology*

Stem cells are a form of cell that can differentiate into other types of cell, in addition they can produce more stem cells through a process of cell-division. Indeed to be classified as a stem cell, certain criteria have been established: "The candidate cell must be capable of asymmetrical cell division, producing an exact multipotent replica cell and an additional progeny cell that can perform a more specialized function" (Sylvester and Longaker, 2004). Due to their capability to differentiate, they play a vital role in tissue development in every part of the body.

There are multiple different types of stem cell that can be distinguished either by their source or their differing capacity to form specialised cell-types. There are broadly three types: embryonic stem cells, adult stem cells and induced pluripotent stem cells. Embryonic stem cells are commonly derived from the inner cell mass of the blastocyst of a pre-implantation embryo (Thomson et al., 1998) and are pluripo-

tent, meaning that they can differentiate to form nearly any cell-type in the body (Schöler, 2016). Adult stem cells, unlike embryonic stem cells, are present in both young and adult humans and mammals. Adult stem cells are multipotent rather than pluripotent meaning that their capacity for differentiation is limited to several distinct cell-types, thus adult stem cells are classified and referred to based on the cells they can produce; neural stem cells for example are capable of differentiating into neuronal and glial cell-types. Lastly induced pluripotent stem cells (iPSCs) are stem cells that can be derived from the reprogramming of mature differentiated cells, these hold great therapeutic promise as they exhibit the morphology and growth properties of embryonic stem cells (Takahashi and Yamanaka, 2006).

Stem cells of any type can be directed to differentiate into specific cells under controlled conditions, these conditions necessarily vary for inducing different cell-types. Factors that have been demonstrated to play a role in directing stem cell differentiation include soluble cues such as cytokines and growth factors, inter-cellular contact, cell-extracellular matrix contacts and physical forces (Clause et al., 2010). Inducing the right combination of these contextual factors within an appropriate time frame can thus cause a stem cell to differentiate into a desired cell-type, however any error in this process can induce a different and unwanted response, for example apoptosis (Clause et al., 2010).

As research into stem cells has progressed there has been much interest in potential biomedical applications. These current areas of research range from the research and treatment of neurodegenerative eye disease (Mead et al., 2015) to the growth of entire organs or organoids for transplantation (Ader and Tanaka, 2014). Stem cells are also of particular benefit to *in vitro* research on tissue function and can be used to replicate diseased tissue through the use of iPSCs from patients, for example those with Parkinson's disease (Devine et al., 2011). Indeed the application of iPSCs to personalised medicine holds great potential (Chun et al., 2011). Practical therapeutic use of stem cell research is still in its infancy in many areas however as there are risks, such as of late tumor formation resulting from inadvertent mutagenesis of stem cells or differentiated products of, that can be difficult to control for (Bubela et al., 2012). Controlling differentiation to ensure all resultant cells are those intended is also an obstacle in present *in vitro* research.

1.2.4.1 *NCAE in Brain Development and Function*

Signaling between cells is of great importance for both development and adult cellular function in multi-cellular organisms. Indeed such signaling can create the context necessary for the differentiation of stem cells into appropriate cell-types. In complex tissues such as the brain inter-cellular communication can be demonstrated through

the role of astrocyte signaling in neuronal development (Allen et al., 2012; Clarke and Barres, 2013), reciprocal signaling between neurons and microglia to regulate neuroimmune response in health and disease (Schafer and Stevens, 2015; Hoarau et al., 2011) also allowing recruitment of stem cells for tissue repair (Hoarau et al., 2011) and the role of microglia in neural circuit development and learning related plasticity (Salter and Stevens, 2017). Such NCAE can occur also between more than two cells, for example a recent study has shown how microglial interaction with astrocytes can cause the latter to express a phenotype toxic to neurons (Liddelow et al., 2017). This inter-cellular signaling can be carried out by a diverse group of factors from cytokines and steroids to growth factors and ligands, the latter for example are used in contact-dependent signaling (Denef, 2014).

As these NCAE are demonstrably important to the development and function of the brain, investigation of the inter-cellular gene expression changes invoked is a means by which we can achieve greater understanding of these processes.

1.3 COMPUTATIONAL METHODS BACKGROUND

In this section I will describe the key computational approaches relevant to the work in this thesis. I will provide an overview of network biology and its common applications, with a focus on the relationship between the structure of biological networks and biological function, before describing how data is integrated and represented through these network methods. I will also summarise the current research on differential network analysis. Subsequently I detail the function of, and approaches to, biological enrichment analysis, where I will provide background on biological pathways and their enrichment before describing enrichment approaches in more detail. Lastly I will overview the application of entropy, as defined in information theory, to research questions in computational biology.

1.3.1 *Network Biology*

Networks, in addition to the methodological techniques that utilise them for analysis, are an essential form of representation within molecular biology. When building a biological network a single source of experimental data, for example a microarray or protein interaction set, is usually used to construct an initial network assembly. This network can either be assembled using only data in the source, if sufficiently rich, or through integration with relevant knowledge from repositories. Common examples of such networks include protein interaction networks, where the networks nodes are distinct proteins with known interactions forming the edges, gene co-expression

networks, where gene nodes are linked by edges weighted by the correlation of their expression, or gene regulatory networks where genes are represented by nodes and edges represent a known regulatory interaction.

Once constructed these networks can be analysed as they are or overlayed with sets of annotation data in order to reveal novel information within the scope of the source dataset through assessment of the network's structure. Indeed it has been long established that the structure of such biological networks can represent and be used to uncover biological function (Pržulj et al., 2004). Annotation of these networks will frequently be applied to label nodes or edges in the network, to adjust information content of labelled nodes or to weight edges. This can be carried out very effectively with a minimum of annotation integration for focuses such as prediction of function (Chua et al., 2006) or disease involvement (Xu and Li, 2006). As a result software packages have been developed to conduct network analysis for a minimum of required data sources (Beisser et al., 2010), for example topographic and over-representation based analysis. In recent years there has been an upward trend in studies using more complex data integration of a higher number of sources and data types (Park et al., 2010; Sun et al., 2014; Gligorijević et al., 2014), an expected development given ever increasing computer capability as well as size and quality of public biological data repositories. In (Park et al., 2010) the authors predict specific biomolecular interactions using genome-wide data: whole-genome networks for yeast, from 3,500 experimental conditions, that include 30 different interaction types. This information was used to investigate pathways in carbon metabolism and cellular transport. In (Sun et al., 2014) systems level biological data sets were integrated to create an a disease network that was used to infer commonality or interaction between diseases. As a last example, in (Gligorijević et al., 2014) the authors integrated information from protein-protein interaction, gene interaction and gene co-expression networks to recreate a significant portion of the Gene Ontology.

These studies have used more complex integration as an approach for asking questions that were previously infeasible to answer given the numbers of biological agents involved or having to require systems level analysis to answer. Indeed a higher degree of integration is becoming standard practice in molecular bioinformatics more generally as the consolidation of greater quantities of evidence naturally allows for an increase in accuracy for predictions and inference (O'Malley and Soyer, 2012). Highly integrative network analysis is widely applicable and so has recently been used to tackle a wide range of biological questions in many areas in bioinformatics from analysis of the human interactome (Pastrello et al., 2014), to the study of macrophage activation (Xue et al., 2014), to large scale connectivity analysis in the mouse brain

(Liska et al., 2015) and finally to topics as diverse as the behaviour and ecology of sharks (Wilson et al., 2015).

However we are also seeing an innovative expansion of how biological networks can be integrated with one another to better understand and infer both topological change within and between established systems and to understand how differences in function can relate to such changes. Network alignment algorithms such as (Pache et al., 2012) can be used to integrate two distinct, but usually related, biological networks together to identify common topology. Aligned areas of these networks can be important for establishing areas of evolutionarily conservation between species, for example the evolutionary relationships between herpes viruses (Kuchaiev and Pržulj, 2011), as well as core areas between condition, temporal samples or species. Cross-species network alignment has been used to infer information about human ageing for example (Faisal et al., 2015). Graphlet methodologies, in which networks topology and change are analysed using small connected sub-networks, also present a vector for network integration (Hayes et al., 2013) and network alignment research (Milenković et al., 2010). Graphlet based network alignment has been used to demonstrate a high level of conservation between the protein-protein interaction networks of human and yeast (Kuchaiev and Pržulj, 2011).

1.3.1.1 *Differential Network Analysis*

Biological network analysis is a reasonably young field and for most of its existence the networks under scrutiny have been built from data of a single experimental condition and have looked to discover dominant features (e.g. protein interactions, gene expression) within the networks of such data. Differential Network Analysis (DNeA) is a broad term for the network analysis methodologies that focus on uncovering the dynamic aspects and changes that occur between different network states, be those condition, sample or temporal in nature. As such this dynamic approach helps to identify which biological agents are directly involved and affected by the changes in experimental condition that we are investigating. As a result DNeA methods encompass many different techniques tailored to the type of network and source data being investigated.

There are broadly four different types of DNeA methodology differentiated through the means and aspects they use to compare different networks. These methods were constructed largely for two sample comparative microarray analysis and so will be described using this context. The first method is an approach known as 'differential wiring' (DW) (Hudson et al., 2009). This approach involves first carrying out a differential expression (DE) analysis on the source data; the difference in co-expression is then calculated, via correlation coefficient, for each pair of genes, with the DW

metric simply defined as the difference in co-expression correlation coefficient for each gene pair between the two samples. A gene's differential connectivity can then be observed as a summary of these metrics between all its connected neighbours. For differentially co-expressed genes the approach then assesses the change in the quantity of significant network edges linked to that gene between the two states and weighting is applied for genes with high 'phenotypic impact', calculated from abundance and DE, and for known regulatory genes. Thus this method uses group and edge specific weighting to identify differential networks, sub-networks that are differentially wired in the two conditions. The second DNeA approach aims instead to identify and compare groups of genes in each network, for example by clustering genes based upon their expression values (Watson, 2006) or using these values to identify minimum spanning trees (Rahmatallah et al., 2014). The gene groups identified in this manner are then compared using a network's summary metrics to elucidate patterns of correlation between samples. The third type of approach focuses on local topological changes (Zhang et al., 2009; Reverter et al., 2006): 'differential connection' in relation to the DE of the genes. The analysis in this approach looks to identify and investigate local network hubs or 'hot spots' where there are both changes in gene expression and in the local network topology between the samples. The last type of DNeA approach has been developed more recently (Ha et al., 2015) and involves splitting analysis into two levels: a local and global component. The local component refers to the identification of differential features in each sample whereas the global component refers to the elucidation of network features that are conserved between samples. This dual-focus approach provides a more in depth analysis of the samples' networks as analysis of the local differential component will not be affected by global consistency between the networks and so observed changes may be revealed to be of higher significance than is perceptible in the previous approaches.

Though the approaches described here were developed mostly for microarray analysis, DNeA is increasingly applied to a wide range of biological data and problems. Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) is a network analysis package, in the style of the second approach described above, that incorporates clustering methodology to identify differentially co-expressed gene groups and can be used to analyse RNA-Seq data. WGCNA clusters genes based upon their co-expression in an attempt to identify functionally similar differentially expressed 'modules' and is widely used within network biology, having been applied to a wide range of biological investigations from studies of cancer (Liu et al., 2017b), to lipid metabolism (Osterhoff et al., 2014) and to identify gene markers for brain ageing (Luo et al., 2018).

It is worth mentioning that concern has been raised (Ideker and Krogan, 2012) around the lack of work regarding statistical issues associated with DNeA: for example cross condition analysis can mitigate the effect of biases that arise as a result of consistent experimental artifacts, whereas independent errors can negatively affect variance when computing network interaction (Ideker and Krogan, 2012). The lack of attention to statistical issues is not necessarily surprising given the reasonably small size of the DNeA field within bioinformatics at large, however the identification and mitigation of such statistical issues within will likely be an important step for algorithmic improvement in the near future.

1.3.2 *Biological Enrichment Analysis*

Enrichment analysis is a common bioinformatics approach to determine the over-representation of biological features in a given sample (Huang et al., 2008). Typically a set of genes of interest, perhaps differentially expressed, will be compared against a knowledge base, the KEGG Pathway Database (Ogata et al., 1999, 1998) or the Gene Ontology (Ashburner et al., 2000) for example, where in such a case the genes will be annotated to biological terms or pathways with each annotated term or pathway then tested for significant enrichment, compared to a background universe of genes, on an individual basis. Significance testing can be carried out with a variety of statistical tests, however the use of Fisher's Exact Test or testing against a hypergeometric distribution are particularly common (Huang et al., 2008).

The product of these analyses is a table of enriched biological features with accompanying p-values, and as such these techniques present a simple powerful means to extract meaning from a group of genes. Biological pathway enrichment for example enriches against 'pathways', groups of genes involved in the same biological process, which is particularly useful for characterising function that may be present in such a group.

Whilst these techniques can prove informative in understanding biological data, interpreting their results if often difficult. The results lists produced are often long and so connections between relevant enriched features can be obscured and hard to notice. The application of multiple testing correction is vital here for ensuring statistical rigour, with hundreds of enrichment tests typically carried out in a single analysis. By ensuring only results that pass this testing are used, the resultant lists are smaller and thus more easily interpreted.

There are several additional issues to be aware of however when conducting enrichment analyses; confounding factors arising from the database used, the background against which the tests are performed and gene pleiotropy. Biological databases are

populated with information derived largely from research literature and in some cases experimentation carried out by laboratories attached to the maintainers, as a result their contents suffer from the positive reporting bias in academic literature (Callaham et al., 1998). In addition, there is a study bias with respect to funding that will also affect database contents; researchers, understandably, tend to prefer using well annotated and well characterised data. As a result, this data gets better characterised through research perpetuating an inequality in annotation through a "rich-getting-richer" pattern (Haynes et al., 2018). Secondly, another key factor to consider is the background set against which enrichment is tested; commonly genes will be tested for over-representation against a background set of genes, this can be all genes known in that species or all genes present in the experimental data. If the background is chosen inappropriately for the analyses, this can lead to the inflation of the resultant test statistics. Finally we have the issue of gene pleiotropy, this refers to the scenario where the expression of a single gene effects more than one resulting and unrelated phenotype. Such genes will necessarily enjoy a better database representation and should be controlled for where possible.

This type of analysis has been used in computational biology for many years, with many tools evolving to use more information to inform their enrichments; for example topGO (Alexa and Rahnenfuhrer, 2010) derives information content from the different types of relationship between terms in the Gene Ontology (e.g. 'is-a', 'part-of') that is used to inform its enrichments, as opposed to ignoring the meaning of these different relationships and simply treating them as edges in a graph. Gene Set Enrichment Analysis (GSEA) tools also provide an alternative approach to that described above, intending to be used with a full list of genes from an experiment rather than just those significantly differentially expressed. GSEA orders genes by expression and calculates an enrichment score that represents the over-representation of genes in a functional set at the top or the bottom of this list, thereby being particularly highly or under-expressed (Subramanian et al., 2005).

1.3.3 *Applications of Information Theoretic Entropy in Bioinformatics*

Entropy, in a general sense, usually refers to a measure of disorder or uncertainty. Entropy in information theory was originally conceived by Shannon in (Shannon, 1948) where he applied the concept to noise in communication. Entropy, as defined by Shannon, is the negative logarithm of the probability mass function of a feature or event, this results in features with high probability carrying less entropy, or 'information', than those with low probability as the occurrence of a lower-probability feature is less anticipated, and therefore more informative. Put simply, the more likely an event

in a given system, the lower its entropy. As such we can understand this entropy to be representative of uncertainty.

Entropy has demonstrated utility through a versatile range of applications in computational biology, from the enhancement of consensus clustering on omics data (Liu et al., 2017a), to the evaluation of substitution rate in viral RNA (Ghasemzadeh et al., 2018) or even to the classification of MRI imaging data (Saritha et al., 2013). Recently entropy has been applied to validation of information derived from gene interaction networks (Wallace et al., 2018).

A notable application is that of the differential network analysis tool 'DiNA' (Gambardella et al., 2013). DiNA applies an entropy based method to tissue-specific gene co-regulation networks in order to determine whether biological pathways are associated with the networks. The entropy probability function here calculates the likelihood of pathway association with each network based upon how many pathway members are co-regulated in that specific network. As a result the entropy will be low when a pathway is well expressed in one network, representing a higher degree of order due to less pathway information being spread across the networks, and high when the pathway's members are present in many networks, with a higher degree of disorder in the pathway's representation and therefore greater uncertainty. DiNA therefore presents an alternative means for achieving pathway enrichment at the network level rather than more commonly used gene set enrichment enrichment analyses whose resolution is the level of a cluster of a simple list of genes. Therefore unlike the latter approach, DiNA is deriving its enrichment from a higher level of information, not simply strength of expression or co-expression, specifically: where the pathway genes are in the clustered network in relation to one another.

1.4 EXPERIMENTAL DATASETS USED IN THE THESIS

There are three experimental datasets used throughout this thesis in order to develop, test and evaluate the novel methods developed. These datasets are described as follows.

1.4.1 *Neuron-Astrocyte Oxidative Stress Dataset (OS)*

The involvement of oxidative stress has been implicated in many acute and chronic neurodegenerative diseases including stroke and Alzheimer's Disease (Halliwell, 2006; Melo et al., 2011; Hardingham and Do, 2016). In order to better understand the effect of oxidative stress on neural cells in these conditions, it is important to investigate

transcriptomic changes that occur when exposing these cells to radical oxygen species (ROS), the active agents that induce cellular damage during oxidative stress.

Astrocytes and neurons were cultured from E17.5 CD1 mouse and E20.5 Sprague Dawley rat embryos; cortices were dissected, enzymatically digested with papain and mechanically dissociated using a 5 ml pipette. Mouse astrocytes were then obtained by growing cells at low density in DMEM containing 10% fetal bovine serum and were passaged twice, using Trypsin. These astrocytes are >99% GFAP positive and <0.1% NeuN positive. At astrocyte days-in-vitro 14, rat neuronal cultures were plated with the anti-mitotic AraC added, which restricts astrocyte numbers to 0.1%. For mixed-species co-cultures, the rat neurons were plated on top of a confluent layer of DIV14 mouse astrocytes. Subsequently, astrocyte monocultures and astrocyte-neuron co-cultures were kept for 8-10 days in Neurobasal-A medium containing B27, but devoid of serum. Neuronal monocultures were kept for 8-10 days in Neurobasal-A medium with AraC. At this juncture, cell cultures were exposed to sub-lethal levels (25 μM) of the reactive oxygen species hydrogen peroxide (H2O2). H2O2 is routinely used to induce oxidative stress and causes a robust induction of genes encoding for antioxidants or stress-response genes. Samples were sequenced both 4 and 24 hours post stimulation, in addition to pre-stimulation controls. Three replicates were taken for each condition.

Henceforth in this thesis, the oxidative stress dataset will be referred to in the following format: OS when referring to the dataset as a whole, the format OSm-[Species initial][Cell-type initial] (e.g. OSm-MN for mouse neuron monoculture) for the the monoculture controls and OScc-[Species initial][Cell-type initial] when referring to cell-type specific data. However only the monocultures from this dataset are used in this thesis. A summary of the oxidative stress dataset can be seen in Table 1.1.

1.4.2 *Neuron-Astrocyte Activity Dependence Dataset (AD)*

This dataset is a two species co-culture of astrocytes and neurons that were exposed to a receptor agonist to stimulate heightened levels of neuronal action potential firing. The experimental procedures were carried out at the lab of my secondary supervisor, Giles Hardingham, as detailed in (Hasel et al., 2017):

Mouse astrocytes were cultured from E17.5 CD1 mouse embryos and rat neurons from E20.5 Sprague Dawley rat embryos; cortices were dissected, enzymatically digested with papain and mechanically dissociated using a 5 ml pipette. Mouse cortical and spinal cord astrocytes were obtained by growing cells at low density in DMEM containing 10% fetal bovine serum and were passaged twice, using Trypsin (both Life Technologies). The resulting astrocytes were >99% GFAP positive and <0.1% NeuN

Organism	*Mus Musculus, Rattus Norvegicus*
Cell Type	Astrocyte, Neuron
Experimental Stimulus	Oxidative Stress
Conditions	Control, 4h, 24h
Replicates	3
Read Length	50bp
Read Type	Paired-end
Sequencing Depth	50 million reads
Monocultures	Mouse Neuron, Mouse Astrocyte

Table 1.1: **Oxidative Stress Dataset Summary**. The key variables and information for the OS dataset. The dataset is comprised of co-cultured mouse astrocytes and rat neurons, in addition to single species mouse monocultures for control, that were subjected to oxidative stress via H2O2.

positive; the rat neuronal cells do contain a small number of non-neuronal cells (principally astrocytes), but since the focus of the study is the mouse astrocytic transcriptome, this was deemed acceptable. At astrocyte days-in-vitro 14, rat neuronal cultures were plated with the anti-mitotic AraC added, which restricts astrocyte numbers to 0.1%. For mixed-species co-cultures, the rat neurons were plated on top of a confluent layer of DIV14 mouse astrocytes. Subsequently, astrocyte monocultures and astrocyte-neuron co-cultures were kept for 8-10 days in Neurobasal-A medium containing B27, but devoid of serum. Neuronal monocultures were kept for 8-10 days in Neurobasal-A medium with AraC. All cultures were used 8 to 10 days after neuron plate down.

The rat neurons were co-cultured with the mouse astrocytes in order to investigate activity dependent gene expression. The co-cultures were transferred into a medium containing tetrodoxin (TTX), a voltage-gated Na^+ channel blocker that prevents action potentials in neurons, for 22 hours to inhibit neuronal firing. Subsequently, TTX was washed from the cells which were then exposed to either the receptor agonist Bicuculline (BiC, 50 mM) to induce network bursting (where periods of rapid action potential spiking are followed by a quiescent periods) or BiC in addition to glutamate transporter inhibitor TBOA (DL-threo-b-Benzyloxyaspartic acid, 50 mM), to increase the half-time of glutamate uptake and thus prolong synaptic activity, for 16 hours, after which the RNA was collected for sequencing. Monoculture control samples for mouse astrocytes, mouse neurons and rat neurons were sampled separately and not for all experimental conditions.

I will thus be using three conditions for this investigation of this dataset: the control where we have TTX dampening neuron activity, leaving only astrocyte activity, the

BiC exposure condition, which triggers a burst of excitatory synaptic activity, and lastly the BiC + TBOA condition, which inhibits glutamate uptake thus exacerbating synaptic activity. There are four replicates for each of these conditions.

Henceforth in this thesis, the activity dependence dataset will be referred to in the following format: AD when referring to the dataset as a whole, ADm-[Species initial][Cell-type initial] when referring to monoculture controls and ADcc-[Species initial][Cell-type initial] when referring to cell-type specific data. A summary of the activity dependence dataset can be seen in Table 1.2.

Organism	*Mus Musculus, Rattus Norvegicus*
Laboratory Strain	CD1 (mouse), Sprague Dawley (rat)
Cell Type	Astrocyte, Neuron
Conditions	BiC, BiC + TBOA, TTX (Control)
Replicates	4
Read Length	50bp
Read Type	Paired-end
Sequencing Depth	150 million reads

Table 1.2: **Activity Dependence Dataset Summary**. The key variables and information for the activity dependence dataset. The original dataset is mixed species with mouse astrocytes and rat neurons. BiC here refers to the receptor agonist bicuculine and TBOA to the glutamate transporter inhibitor DL-threo-b-Benzyloxyaspartic acid.

1.4.3 *Neuron-Astrocyte-Microglia Three Species Dataset (3Scc)*

This dataset is a three species RNA-Seq experiment containing neuron, astrocyte and microglial cell-types. The neurons are derived from mouse stem cells, the astrocytes from human and the microglia from rat. The experimental procedures were, once again, carried out at the lab of my secondary supervisor, Giles Hardingham, as detailed in (Qiu et al., 2018):

Human primary astrocytes were cultured for 1-3 days until they reached 80-90% confluency. Mouse neurons derived from embryos were then plated atop the astrocytes, these were then incubated for 3-4 days. Four wells per condition and per co-culture type were used. Rat microglia were extracted from neonatal pups and cultured separately for 12 days before being plated with the astrocyte-neuron co-culture for the three species co-culture conditions. Immunostaining was carried out to confirm samples were free of contamination, but was not carried out for each replicate. Three days after plating down the microglia, the co-cultures were challenged with the

inflammatory simulus, 500ng Lipopolysaccharide (LPS), which is an activator of microglia and triggers toll-like receptor 4 signaling. RNA was collected for sequencing 24 hours after the addition of LPS. This methodology allows for microglia-dependent transcriptional changes to be tracked in both neuronal and astrocytic cell types simultaneously, both with and without the LPS-induced response of the microglia.

From this experiment I will thus be using four conditions to build the gene co-expression network for this investigation: a control where we have an unchallenged co-culture of neurons and astrocytes, a second control where we have an unchallenged co-culture of neurons, astrocytes and microglia, an exposure of the neuron and astrocyte co-culture to 500ng of LPS, and an exposure of the neuron, astrocyte and microglia co-culture to 500ng LPS.

Henceforth in this thesis, the three species dataset will be referred to in the following format: 3Scc when referring to the dataset or 3Scc-[Species initial][Cell-type initial] when referring to a cell-type specific portion of the dataset. A summary of the three species dataset can be seen in Table 1.3.

Organism	*Mus Musculus, Rattus Norvegicus, Homo Sapiens*
Cell Type	Neuron, Microglia, Astrocyte
Conditions	Control, 500ng LPS (Both conditions with & without microglia)
Replicates	3
Read Length	75bp
Read Type	Paired-end
Sequencing Depth	Conditions without microglia: 100 million reads, Conditions with microglia: 220 million reads

Table 1.3: **Three Species Dataset Summary**. The key variables and information for the three species dataset. This dataset is mixed species with mouse neurons, human astrocytes and rat microglia. 'LPS' here refers to 'Lipopolysaccharide' which is introduced to the cultures to trigger an immune response from the microglia.

1.5 CONTRIBUTIONS

The scientific contributions laid out in this thesis can be summarised in the following:

1. The development of a computational methodology, known as 'Sargasso', for the *in silico* separation of mixed-species RNA-Seq datasets. This tool is intended for use with datasets wherein distinct cell-types are cultured from distinct species where a separation by species therefore represents a partitioning of the data

by cell-type. Sargasso provides several strategies for optimising species separation, is publicly available as a Python package and has been published by peer-reviewed journals for its use in the investigation of non-cell-autonomous effects (Hasel et al., 2017; Qiu et al., 2018).

2. The development of a computational pipeline, known as 'Pathway Entropy', which improves upon previous work (Gambardella et al., 2013) to provide an application of information theoretic entropy to gene co-expression networks for the purpose of pathway enrichment. This method provides more accurate and interpretable results when applied to clustered gene co-expression networks than commonly used hypergeometric methods and that of (Gambardella et al., 2013).

3. The evaluation of the performance and robustness of these novel methods when applied to the investigation of non-cell-autonomous effects with both methods having been subsequently applied to separate two and three species mixed-species RNA-Seq datasets.

1.6 ORGANISATION OF THE THESIS

In the second chapter I focus upon the design and development of the novel 'Sargasso' methodology, a tool to enable the *in silico* separation of mixed-species RNA-Seq data. I describe the testing of this method on simulated and experimental RNA-Seq datasets and assess its performance with regard to a number of key factors: accuracy, data loss and impact on downstream analysis.

The third chapter describes the design and development of the novel 'Pathway Entropy' methodology, an approach that applies information theoretic entropy to clustered gene co-expression networks in order to determine the involvement of biological pathways. This tool is an improvement upon a previous method, 'DiNA', against which it is evaluated in addition to other commonly used tools for pathway enrichment. Its performance is evaluated with regard to the contribution of information, namely edge weights and pathway connectivity, that its variant implementations provide.

In the fourth chapter I detail the application of the novel methods from chapters two and three to two mixed-species RNA-Seq datasets. Their performances are examined in terms of consistency and accuracy and the results are placed in context with reference to previously published relevant biological results.

In the final chapter, I reflect upon the contributions of this body of work and draw conclusions from the results presented in light of the study of non-cell-autonomous

effects and network biology. I discuss the limitations of the work presented in the thesis and how future work could be conducted to extend it.

Lastly, in the appendices I include tables of raw results used to generate the graphs in the thesis for reference.

1.7 OUTLINE OF INDIVIDUAL CONTRIBUTIONS TO PROJECTS IN THIS THESIS

This thesis describes the development of two principal software development projects: Sargasso in Chapter 2 and Pathway Entropy in Chapter 3. In this section I will detail the contributions of myself and others to these projects to make clear my role in their development. The application of these methods in Chapter 4 was entirely my own undertaking.

The development of Sargasso was a collaborative undertaking by myself and my supervisor Owen Dando. Owen had performed the initial feasibility study for genome separability and I began my PhD to develop the tool in earnest. For the length of the tool's development I was thus working on this project full-time along with Owen, who was working on it alongside his other responsibilities. It is both difficult and impractical to disentangle individual contribution for most aspects of the development process, as is common in collaborative software development, as each stage of the process was conceived through a constant dialogue of discussion and collaborative research. Subsequent implementation and editing were carried out by both of us in order to best identify bugs, with revision and re-implementation of code portions a frequent occurrence. Thus even for portions that may have begun as the work of one of us, the finished code is necessarily the result of this collaborative process.

There are however, two areas where individual contribution is certain: the development and implementation of Sargasso's assignment algorithm was exclusively my work and the compilation of the finished Sargasso code into its present format, a python package, was the exclusive work of Owen Dando. After I started work on the Pathway Entropy method, our colleague Xin He lent his efforts to aid continued maintenance and development of the package, thus making the most recent code a three-way collaboration. In Chapter 2, there are several figures included that were plotted from our data by Owen for our future methods paper. As these figures were not plotted by my own hand I have thus labelled them as belonging to the work of our paper currently in production, using the label: 'Heron, Dando and Simpson (Forthcoming)', in order to make this explicit.

The development of the Pathway entropy tool in Chapter 3 was not a collaborative project, all coding and development were an exclusive undertaking by myself.

2

SARGASSO

2.1 MOTIVATION

The mechanisms which underlie biological systems, cell-autonomous and not, are important for our understanding of cells and cellular functions as well as of cancer and neurological disease. Presently much knowledge here is restricted to cell-autonomous settings as it is simpler to study the functioning of a single cell-type or monoculture. Studies investigating non-cell-autonomous mechanisms commonly measure gene expression levels in different cell types, but in order to partition these populations the separation procedures available are physically invasive, damaging the cell cultures and bias the very gene expression levels extracted for study. Indeed it has been shown that physical separation techniques for cultures of mixed cell-types trigger stress and apoptosis related genes, introducing noise which obscures identification of genes of interest and introducing bias (Okaty et al., 2011b,a; Hasel et al., 2017).

Whilst the negative effects of physical separation techniques have long been accepted as unavoidable in the field, they present serious problems for downstream analysis. The addition of noise can mask the detection of genes whose differential expression is small between conditions and the bias toward cell stress and death genes can result in misleading conclusions and hamper attempts to analyse the genetic data at the level of systems or pathways.

This relative difficulty in the study of non-cell-autonomous effects has resulted in the present situation where we are reasonably confident about the function and action of many cellular types, on a cell-autonomous level, however have less of an idea about how these individually characterised types interact, either *in vitro* or *in vivo*, and how these interactions shape the function and action of the cells they interacts with. Thus an increased ability to study these effects would substantially improve our understanding and characterisation of even well understood cell types.

A key area that would benefit from such study is the *in vitro* culturing of stem cells, particularly for neural cultures. Presently, due to a lack of knowledge surrounding neural cell maturation, neural monocultures do not fully develop the morphology of *in vivo* neurons and frequently include undeveloped stem cells (Woodbury et al., 2000; Roy et al., 2006). These undeveloped cells pose a tumour risk if used in xenograft implantation and bias research efforts *in vitro*. There is also the risk that stem cells

may develop and mature but into cells other than those desired and in turn bias results collected from the culture.

The quality of reference genomes and genome annotations for established laboratory species continues to improve with research year on year. With such higher quality now available, we posit that if a different species was used for each cell type of study, we should now be able to attribute genomic data *in silico* through cross mapping RNA reads to the relevant genomes. In such a scenario, issues of cell stress, cell death and expression bias from separation procedures would no longer apply, facilitating easier study of non-cell-autonomous processes and effects. The use of different species may of course introduce different biases however. With regard to this point, the work of my collaborators has demonstrated that the use of different species, mouse astrocytes with rat neurons and vice versa, induces comparable effects on astrocytic marker genes (Hasel et al., 2017). Species difference was also shown not to affect activity dependent expression in a subset of genes that were analysed more closely (Hasel et al., 2017).

In this chapter, we present a novel approach that alleviates the issues introduced by physical separation in gene expression quantification using RNA-seq by conducting *in silico* sequence separation between different sources. Our method uses data derived from a novel experimental paradigm, mixed species co-culturing (e.g. two cell types each belonging to closely related species, or indeed different strains of the same species), and maps the resultant RNA-Seq differentially between the two genomes. The mappings for each read are then assessed by alignment quality factors and assigned to one source genome or another according to user specified selection criteria. Closely related species, such as mouse and rat, are used in order to minimise any biological differences that may arise from evolutionary divergence and in order to achieve as close as possible a representation of *in vivo* cellular behaviour.

Our method is assessed for both performance and recall in addition to species choice. Any effect on downstream analysis, such as differential expression, is also investigated. Our novel approach allows us to investigate differential transcriptomic effects between cell types and our evaluation demonstrates this in the context of two neural cell types: neurons and astrocytes. While astrocytes have commonly been understood to act as support cells or 'neural glue', previous research (Johnson et al., 2007) has challenged this view and demonstrated a positive effect of astrocytes when co-cultured with neurons on cell development, morphology and synaptogenesis. Our evaluative findings contribute to the better characterisation of this cellular relationship.

This project was a collaborative undertaking by members of my group with the principal contributors being myself, Owen Dando and Xin He under the guidance of

my supervisors Ian Simpson and Giles Hardingham. The Sargasso tool continues to be maintained and developed by the group.

2.2 DESIGN & IMPLEMENTATION

In this section I will describe the details of the Sargasso methodology we have developed and how it has been applied to experimental data. I will detail the tools that form part of its pipeline as well as justification of approaches taken before elaborating on the separation methodology itself.

2.2.1 *Approach*

2.2.1.1 *Language*

This project was coded primarily in the Python programming language due to familiarity with the language and availability of established libraries for computational biology. This was particularly important for the interfacing of file processing tools, such as `Sambamba` (Tarasov et al., 2015). For some components, Linux shell scripting was used to most efficiently make use of commands for executing tools manipulating results files; this could have been coded through Python however we deemed such wrapping unnecessary.

2.2.1.2 *Platform*

The Sargasso package has been designed on systems running Linux distributions and it requires the shell environment, common to all Unix-like operating systems with substitutes available on Windows systems, in order to be executed. The tools and libraries that Sargasso requires are not platform specific and so do not strictly limit its usage to a Linux platform.

2.2.1.3 *Intended Use Cases*

Sargasso is intended to be used to separate, by species, mixed species RNA-Seq datasets in which the species present are known. These datasets can be the product of laboratory testing or of simulated *in silico* origin. In order to achieve approximate biological compatibility between cells, it is recommended to use closely related species or species in whom the biological function of interest is well conserved. Distant species' cells may reject one another or behave differently, compromising the purpose of the investigation.

To ensure accuracy of findings it is important to use species whose reference genome and annotation do not contain large gaps of missing information as this could jeopardise sequence read assignment accuracy. Whilst Sargasso was developed

initially for separating datasets containing data from two species, the separation of datasets with any number of species is now possible.

Whilst we have not tested all possible combinations of model organisms, we have applied Sargasso to the separation of mammalian species including: mouse, rat, human, macaque, chimp and cow, in addition to chicken, zebrafish and frog. For species of particularly closer genetic distance, such as human and chimp, we do see a reasonable read loss due to genome similarity, however a separation is still possible. Further detail on the separation two species can be seen in the supplementary material of (Qiu et al., 2018). As such for the intended use case with regard to genetic distance of species, our recommendation would be those with a distance of approximately 10 million years.

2.2.2 *Separation Mechanism*

In order to disambiguate the reads of different species *in silico*, we must carefully identify a parameter or series of parameters capable of distinguishing the parent organism of each read. The raw RNA-Seq data itself does not contain such parameters and so we must derive them from elsewhere. An obvious candidate for such parameters would be those produced through genome alignment.

2.2.2.1 *Assessment of Key Variables*

The read aligner `STAR` (Dobin et al., 2013) was used to map the RNA-Seq reads to the genome of each species in the data. The resultant alignments are output in BAM format and contain a number of variables describing the quality of each alignment. We investigated those we deemed most discriminating and these are detailed further below. As RNA-Seq reads are sequenced a different lengths, commonly denominators of 50bp, variation in these parameters was investigated by proportionately weighting the threshold for length for any increase above 50bp reads.

Alignment Score

Each alignment has an alignment score (AS) generated for it that represents how similar the read is to the reference. AS increases with each matched base and decreases with the number of mismatches and for any gaps. The exact scoring depends on the scoring matrix used to score the alignment of bases in a sequence alignment.

This variable seemed from the outset a logical discriminator given that it takes other alignment metrics into account. However, upon testing we discovered that the score is not comparable between mappings to different genomes. This is due to the

manner in which introns are scored. As the AS incorporates the length of an intron, whose length even in well conserved regions may differ between species, reads mapped across splice junctions will thus have a AS derived from species specific information.

As such if we wish to use AS as a discriminating variable we will have to discard all reads mapped with over splice junctions as we cannot compare these alignments. As this presents a flaw that could lead to incorrect assignment, we looked instead into the possibility of using several of the variables included in the calculation of AS which are not species specific.

Mismatches

The number of mismatches in an alignment is the number of bases for which there is disagreement between the read and the reference. Single mismatches could represent sequencing errors or point mutations, however any increase beyond this point casts doubt over an alignment's plausibility, especially if a read is aligned multiple times to a genome as this further increases uncertainty.

As the number of mismatches is a good indicator of alignment quality, we chose to use it as a discriminator. On its own however it is not enough to reliably capture the full context of the alignment quality so we use other discriminators in addition.

Multiple Mappings

The number of times a read maps to a genome can be used as a good indicator of alignment quality. Reads that are mapped to multiple locations could have suffered errors during sequencing or indicate simply partial conservation or similarity of a read from one species upon the genome of another. For whatever reason they may be present, they introduce uncertainty when it comes to assessing assignment. Thus in conjunction with other discriminators, such as the number of mismatches, the number of multiple mappings is a useful comparator for mapping quality.

Should we want to be stringent in our investigations, throwing out reads that map multiple times to each species is a good way of minimising uncertainty in assignment and thus reduce possible false positives.

Minimum Match

The 'minimum match' variable informs us of how much of a read is aligned to the genome. A read may seemingly align perfectly with no mismatches or multiple

mappings, however half of it could be 'clipped', or removed, from the alignment. Clipping is taken into account by AS, however cannot be assessed from just from the use of the mismatch and multiple mapping variables. As such we will also make use of the minimum match as a discriminating variable.

Spliced Alignments

Finally, for spliced alignments, we recognise that read alignments that overhang on one exon by a single base or two, but are otherwise identical for each species they are mapped to, are potentially spurious as read quality is known to worse at the end of reads (Fuller et al., 2009). Thus we require that the minimum length of sequence aligned to any one exon must be at least five bases; in testing, we found that alignments with exon overhangs shorter than this length were frequently the source of incorrect species assignment.

2.2.2.2 *Proposed Mechanism*

As we have three proposed variables for use in determining a read's species of origin, we need a mechanism to determine under which circumstances we assign a read, or not. In the case where alignment to one species is perfect but to the other is poor, this is trivial; however if we have non-optimal alignments to both species then we need to determine how these discriminating variables are prioritised.

In such a case, and with the presumption that the values for these variables fall within acceptable parameters set by the user, then the alignments are directly compared as follows:

1. If the minimum number of mismatches with respect to the reference genome amongst all mappings in the alignment set is fewer for one species than for the other, the read is assigned to the former species.

2. If the minimum number of mismatches is the same for each set of alignments, the length of alignments is examined. If the mappings to one species' genome are of full read length, whereas the other set contains alignments that are clipped, the read is assigned to the species with full length mappings.

3. If all the alignments to both species' genomes are full length (or both sets of mappings contain alignments which are clipped), then the number of multi-mappings of the read to each species' genome are compared; the species with fewer multi-mappings is chosen at the likely species of origin.

4. If the number of multi-mappings is the same for each alignment set, the read is declared ambiguous and unassigned.

For cases where we are allowing multimaps, the first two steps compare the mapping with the minimum number of mismatches and the maximum alignment length respectively.

2.2.3 *The Sargasso Pipeline*

2.2.3.1 *Overview*

The Sargasso methodology is structured as a pipeline. This pipeline includes both pre-existing tools and novel code and is displayed in full in Figure 2.1.

Given an input set of FASTQ files for a number of samples, containing mixed-species RNA-seq read data, the Sargasso pipeline partitions reads according to their likely species of origin, and outputs per-sample BAM files containing the mapping of these reads to the respective genomes. This separation of species data is performed in a number of stages, of which the most important steps are:

- ***Mapping***: Raw mixed-species RNA-seq data is mapped to each species' genome using an efficient splice-aware read aligner: we use `STAR` for this purpose.
- ***Sorting***: Mapped RNA-seq reads are sorted in name order so that, subsequently, the mappings for each read (or each read pair, in the case of paired-end reads) to each genome can be assessed together, in direct comparison. For this we use `Sambamba`.
- ***Assignment***: On the basis of its genome mappings, each RNA-seq read or read pair is assigned to its correct species of origin or discarded as ambiguous if its species of origin cannot be determined.

The mapping and assignment steps are examined in detail below, the sorting being a trivial process. Also detailed are the downstream steps of read counting (e.g. using `featureCounts` (Liao et al., 2014), all command line tools henceforth are displayed in fixed-width font) and differential analysis (e.g. using DESeq2 (Love et al., 2014)) described in Figure 2.1. However, we first note that while the Sargasso pipeline allows the user fine control at each stage, in normal usage the pipeline up until read summarisation can be executed automatically with a single command.

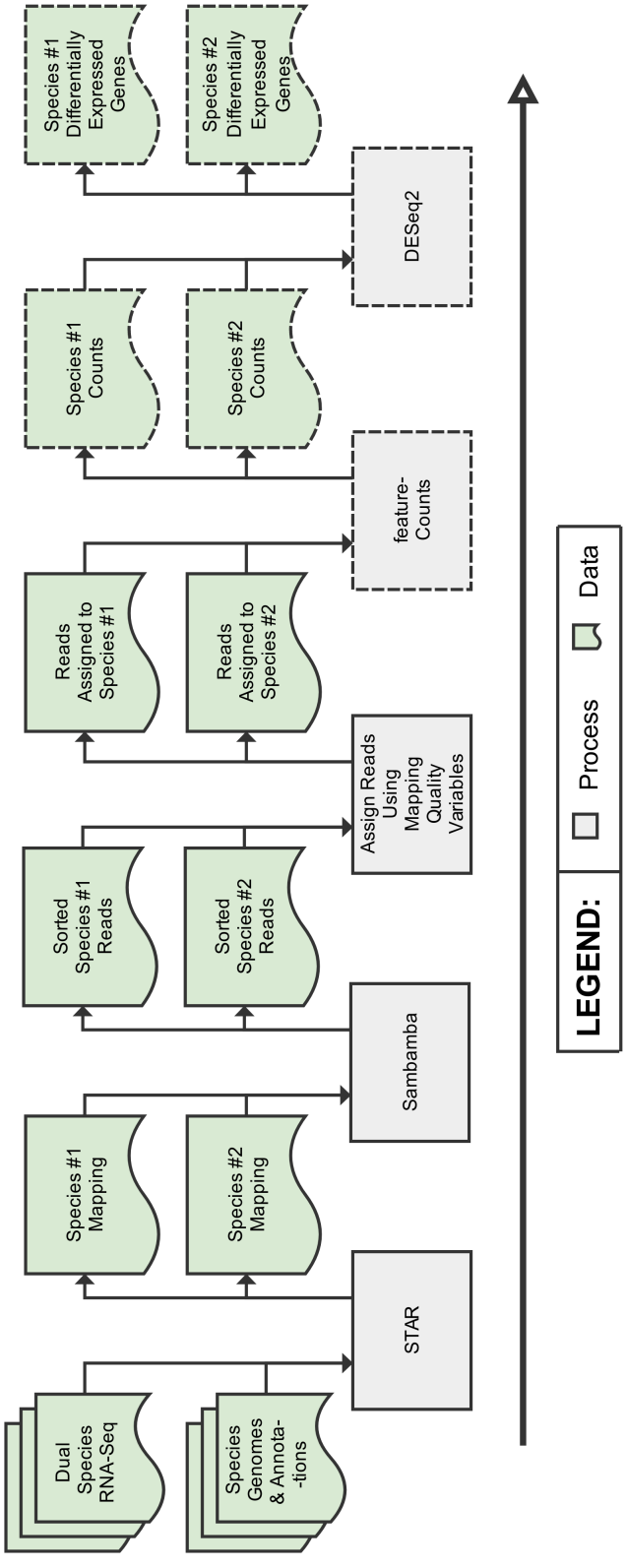


Figure 2.1: **Sargasso Pipeline**. An overview of the Sargasso pipeline's core functions (in grey) in a two species scenario and any key data (in green) that they produce or require as input. `featureCounts` (Liao et al., 2014) and DESeq2 (Love et al., 2014) are included here as an example of downstream analysis, indicated by the dashed outline.

2.2.3.2 *In Detail*

Read Alignment with `STAR`

While we chose to use the `STAR` read aligner both for its speed and the accuracy of alignments produced (Engström et al., 2013), it has since continued to prove a robust and reliable approach for RNA-Seq alignment (Ballouz et al., 2018; Baruzzo et al., 2017) and continues to be chosen for its performance in recent literature (Raplee et al., 2019; Lachmann et al., 2018). However should users wish, it would be possible to substitute another mapping tool as subsequent analysis steps require only BAM formatted files containing the aligned reads.

Reads are mapped allowing alignments to multiple locations (`-outFilterMultimapNmax 10000`); however only those mappings with an AS equal to the maximum are retained (`-outFilterMultimapScoreRange 0`). As AS here is not used for comparison, but rather to ensure quality with respect to genome mapping, species inconsistency in intron sizes are not an issue. These parameters, or equivalent, would need to be applied to ensure consistency of Sargasso's performance should users decide to use an alternative read aligner.

Subsequent to alignment, the BAM files produced by mapping reads to each species' genome are sorted into name order using the `Sambamba` alignment processing tool.

Species Assignment

Once reads have been sorted into name order, the mappings to both genomes for each read or read pair can be examined in turn to determine the species of origin for each read. Firstly, if a read has alignments to one species' genome, but none to the other, it is provisionally assigned to that species. While it would be tempting to make this assignment definitive, it must be taken into account that the reference genomes to which reads are mapped may be of different qualities. This may be true both in the extent and accuracy of their raw sequence and in the standard of their annotation. Thus the lack of alignments to a species' genome does not necessarily preclude a read belonging to that species.

To mitigate this issue, for a provisionally assigned read, mappings are required to satisfy a number of user-defined thresholds:

- ***Number of multi-maps***: A given read must have *at most* a certain number of alignments to the putative species of origin's genome. Multiple mappings of low quality to one species' genome may be indicative that a species may not be

the origin of a read, but may also indicate that its locus is missing or incorrectly sequenced in the reference genome.

- ***Maximum number of mismatches***: Read alignments must not exceed a certain number of bases, with respect to their length, being mismatched with respect to the reference genome. This is implemented as $n/50bp$ so as not to penalise shorter reads, where n is the supplied maximum number of mismatches.

- ***Minimum length of alignment***: All read alignments must exceed a certain proportion of read length, including both correct matches and mismatches with respect to the reference genome. That is, a minimum proportion of clipping of the read is tolerated. This parameter is also scaled for read length in the same manner as mismatches: $m/50bp$ where m is the supplied length

- ***Spliced alignments***: For spliced alignments, the minimum length of sequence aligned to any one exon (by default) must be at least five bases. Alignments over splice junctions for a mere base or two are unreliable where this is the only portion differentially mapped due to read mapping becoming less reliable at the end of a read (Fuller et al., 2009). As such a minimum threshold will reduce the risk of false positives.

For a given read, if its alignment to a genome satisfies these criteria and if it is only mapped to one genome, it is assigned to that species. The exact thresholds used can be chosen to balance between the precision of assignment of reads to their correct species of origin and the recall of the maximum number of reads. These choices of filtering strategy are discussed later in this section.

If a read maps to more than one species involved, we compare the criteria listed above for each species. Each set of alignments is tested against the thresholds detailed above. If the criteria are violated then its mapping to that species is disqualified; if this is the case for all species then the read is rejected and left unassigned.

In the final case, that the mappings to more than one species' genomes satisfy all thresholds, then these sets of mappings are directly compared by the algorithm previously described in Section 2.2.2.2.

Filtering Strategies

Through the choice of the particular thresholds that RNA-seq reads must satisfy in order to be assigned to their true genomes of origin, different filtering strategies may be adopted. Such strategies provide a particular balance between precision and recall (or sensitivity and specificity). For example, demanding a high specificity, that

the number of reads mis-assigned to the wrong species is minimised, may result in a reduction in the total number of reads correctly assigned. Similarly, the requirement for high sensitivity, that the total number of reads assigned is maximised, may result in a concomitant increase in the percentage of reads incorrectly attributed.

The particular strategy desired by the user can be finely adjusted through parameters which control the thresholds governing the number of multi-maps, number of mismatches, and length of alignments that reads are required to satisfy. Below, we demonstrate that Sargasso shows good performance across a wide range of parameter values, with high proportions of the total read set correctly assigned and low numbers of reads incorrectly attributed. However, we also show that a particular choice of parameter values provides excellent behaviour in a wide variety of situations, with both high precision and recall, whichever the species of origin of the mixed-species RNA-seq data. For ease of use, the combination of parameter values that make up this filtering strategy can be set through a single command-line option.

However, for cases where it is of particular importance that the filtering strategy is as conservative as possible, that minimising the number of reads mis-assigned takes foremost priority, we demonstrate below that such a conservative approach still achieves high sensitivity. Again, for convenience, the parameter values settings for this filtering strategy can be set through a single command-line option.

Read Summary Tool

A tool for read summarisation, whilst not a core step in the Sargasso pipeline, is necessary for feature-wise downstream analysis. As such I will briefly discuss my choice of tool in the development of Sargasso and the subsequent analysis of species separated data.

When prototyping the Sargasso pipeline, I initially used the tool `HTSeq` (Anders et al., 2015) as it is an established standard for read summary that interfaces directly with the commonly used DESeq2 differential expression tool. It is known to be slow as it is coded in Python and not optimised for speed. As speed of processing is important for downstream analysis given the higher quantity of data that is output by Sargasso, an alternative tool was sought.

`featureCounts` is a more recently developed command line tool that is optimised for speed. Through comparative analysis it has been demonstrated (Liao et al., 2014) that whilst the overwhelming majority of reads were assigned to the same genes by both tools, `HTSeq` assigned 67 reads (less than 0.001% of total read count) to genes that `featureCounts` did not, whereas `featureCounts` assigned 27102 reads to genes that

`HTSeq` did not. Given that there is a small performance gap in favour of it in addition to a faster run time, `featureCounts` was used for purposes of downstream analysis.

`featureCounts` was executed using default parameters.

Downstream Analysis - Differential Expression

For the purposes of testing and evaluating changes to the core Sargasso pipeline, the R package DESeq2 was used to measure the differential expression of genes in the species separated RNA-Seq datasets. DESeq2 was chosen as it is well established and we were more familiar with its use than for the similarly well established tool edgeR, though the sensitivity for both tools is reportedly similar (Love et al., 2014). This tool was executed using default parameters.

2.3 RESULTS

2.3.1 *Application to Simulated Data*

To test the utility and function of Sargasso, we first chose to do so using simulated data as we had not yet the experimental data to work with. We simulated RNA-Seq reads from two species that have a small genetic distance between them in order to analyse Sargasso's capability to correctly attribute RNA-Seq reads to their species of origin. Rat (*Rattus Norvegicus*) and Mouse (*Mus Musculus*) were chosen as they are of close genetic similarity and are both also primary models of scientific study, thus ensuring higher quality reference genomes. The ability to separate them also demonstrates a use to those conducting both behavioural and transgenic research as these species represent the primary animal models for these modes of study respectively.

Rat reads were simulated error free using Flux Simulator (Griebel et al., 2012), an *in silico* read simulator for RNA-Seq data, at both short (50bp) and long (150bp) read lengths. Sequencing error typical of Illumina sequencing, taken from supplementary Table S2 in (Jia et al., 2013), was then imposed manually on these sets using quality profiles derived from experimental rat RNA-Seq datasets of such lengths. These were taken from GEO (Edgar et al., 2002): GSM1708531 (50bp) and GSM1630458 (150bp). Error was incorporated through weighted random substitution, weighted by instance in the profile for the position of each base, using the proportions from (Jia et al., 2013). Error was not simulated using Flux Simulator directly as it contains error profiles for 35bp and 70bp read lengths only. Reads for the mouse were not simulated as by using data from a single species we could accurately determine the efficacy of the read assignment.

The simulated datasets were supplied to Sargasso which mapped them to the genomes of both rat and mouse. The read mappings were then assigned to the species using conservative values for the key assignment variables; no mismatches, no multiple mappings and only full length mappings. This was done both to promote precision over recall and to establish a baseline performance for separation with Sargasso. The results for both read lengths, with and without simulated sequencing error, can be seen in Table 2.1. Whilst a higher number of overall reads for these datasets would have better approximated the RNA-Seq datasets we work with, difficulty with Flux Simulator prohibited sets larger than these.

	Rat				Mouse					
	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous	Precision	Recall
50bp	131508	124191 (**94.4%**)	5740 (**4.4%**)	1577 (**1.2%**)	60603	6 (**0.0%**)	59020 (**97.4%**)	1577 (**2.6%**)	0.9999517	0.9443608
50bp (Error)	131266	105380 (**80.3%**)	24549 (**18.7%**)	1337 (**1.0%**)	59493	7 (**0.0%**)	58149 (**97.7%**)	1337 (**2.3%**)	0.9999336	0.8040224
150bp	43885	42065 (**95.9%**)	1553 (**3.5%**)	267 (**0.6%**)	18830	0 (**0.0%**)	18563 (**98.6%**)	267 (**1.4%**)	1.0000000	0.958528
150bp (Error)	43709	26713 (**61.1%**)	16826 (**38.5%**)	170 (**0.4%**)	18051	0 (**0.0%**)	17881 (**99.1%**)	170 (**0.9%**)	1.0000000	0.6111556

Table 2.1: **Read Assignment for Simulated Rat Reads.** This table contains the quantity of reads assigned to, rejected from and unassigned to either species for simulated rat reads of 50bp and 150bp length both with and without sequence error. Where 'Assigned' refers to reads that have been assigned to a species having determined it to be the species of origin, 'Rejected' being the number of reads that were mapped to the species but not assigned as a result of either mapping better to another species or exceeding the threshold of one or more quality variables. Lastly 'Ambiguous' refers to the number of reads that were not assigned as a result of mapping equally well, or poorly, to all species. 'Total Reads' is the sum of the assigned, rejected and ambiguous reads for each species, these figures are only counted for reads that aligned to the genome of a species and so the total reads for each species will be different.

Sargasso's read assignment for the simulated reads can be seen here to be remarkably effective in terms of precision, the proportion of reads assigned to the correct species, as we can see that the number of reads incorrectly assigned to mouse is low for all samples and thus that of all reads assigned to a species, the great majority have been correctly assigned to the rat. We see Sargasso's performance expectedly lower for recall, the proportion of rat reads assigned to rat, where due to the stringent thresholds of the conservative filtering approach we see many rat reads be fail to be assigned to rat, mainly through rejection rather than ambiguity, and thus contribute to a lower proportion of all reads being correctly assigned.

2.3.2 *Assignment Accuracy*

In order to improve the recall we varied the key assignment variables both independently, where one variable was altered with the remaining two consistently at conservative threshold, and in conjunction with one another to explore the degree to which they affect read assignment and also that to which they demonstrate independence. This was carried out using the same simulated rat reads and separating against rat and mouse genomes. The results of the independent criteria variation can be seen in Figure 2.2 A,B and C, whilst the results of the variation of the criteria in conjunction can be seen in Figure 2.3. Change in precision and recall here is represented using the F1 score which is calculated as the harmonic mean of the two. As we have previously demonstrated a high precision any marked increase in F1 will likely be due to an increase in recall.

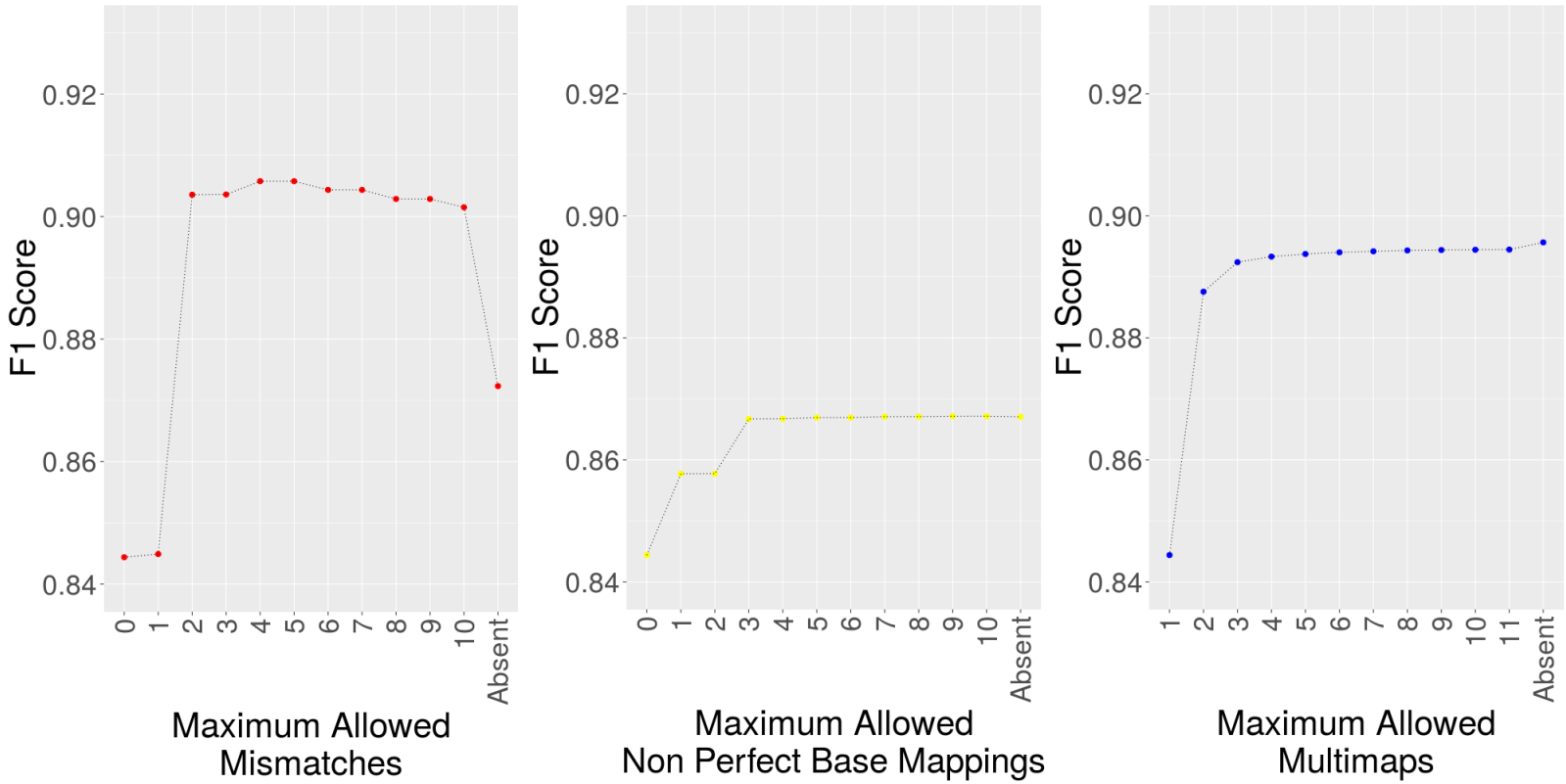


Figure 2.2: **Change in F1 Score for Independent Variation of Assignment Criteria.** Independent variation of assignment criteria has a demonstrable effect on the F1 for Sargasso's assignment, with the number of mismatches (A) demonstrating the greatest impact. Whilst increasing the number of bases that can be spliced out of the alignment (B) also has a demonstrable positive impace on F1. Interestingly not using the multimap criteria results in a higher F1 (C) than achieved by relaxing the threshold. These results were generated for a simulated rat RNA-Seq sample of 50bp read length. The x axis represent thresholds for these criteria, with 'Absent' referring to assignment with only the other two criteria at conservative threshold. The y axis displays the F1 Score.

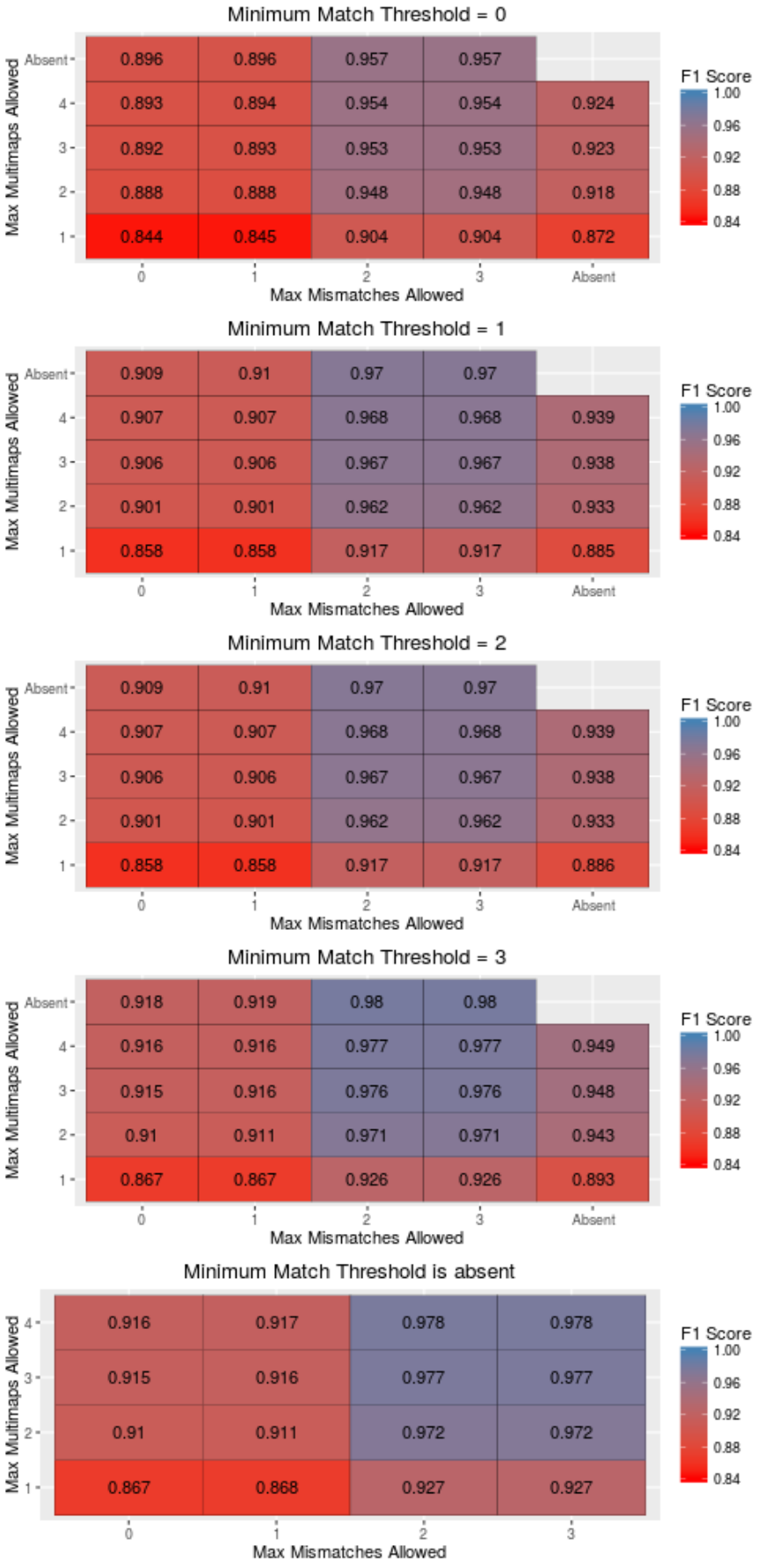


Figure 2.3: **Change in F1 Score for Co-variation of Assignment Variables.** Varying the three key assignment variables in conjunction; mismatches (X-axis), multi-maps (Y-axis) and minimum match (variation per plot), demonstrates that the highest F1 score is achieved when assignment minimum match and mismatch thresholds are relaxed in the absence of thresholding for multiple mappings. The F1 score is calculated as the harmonic mean of precision and recall (F1 Score) using simulated rat RNA-Seq of 50bp read length.

Figure 2.2 shows that relaxing the thresholds for the key variables allows for a significant increase in F1. Yet most of this increase is accounted for in the first or initial few steps of threshold decrease, with subsequent relaxations yielding progressively fewer returns and necessarily increasing potential for incorrect mapping.

Surprisingly, the removal of the multi-mapping variable results in a marginal increase in F1. Comparatively the removal of the mismatch criteria, whilst an improvement from its conservative values, is of significant detriment to the F1 from its more relaxed thresholds. The removal of the minimum match criteria produces an F1 only marginally lower than the F1 of the most relaxed threshold tested on this graph. As a result of this, for the purpose of optimisation we will consider only the minimum match and mismatch criteria.

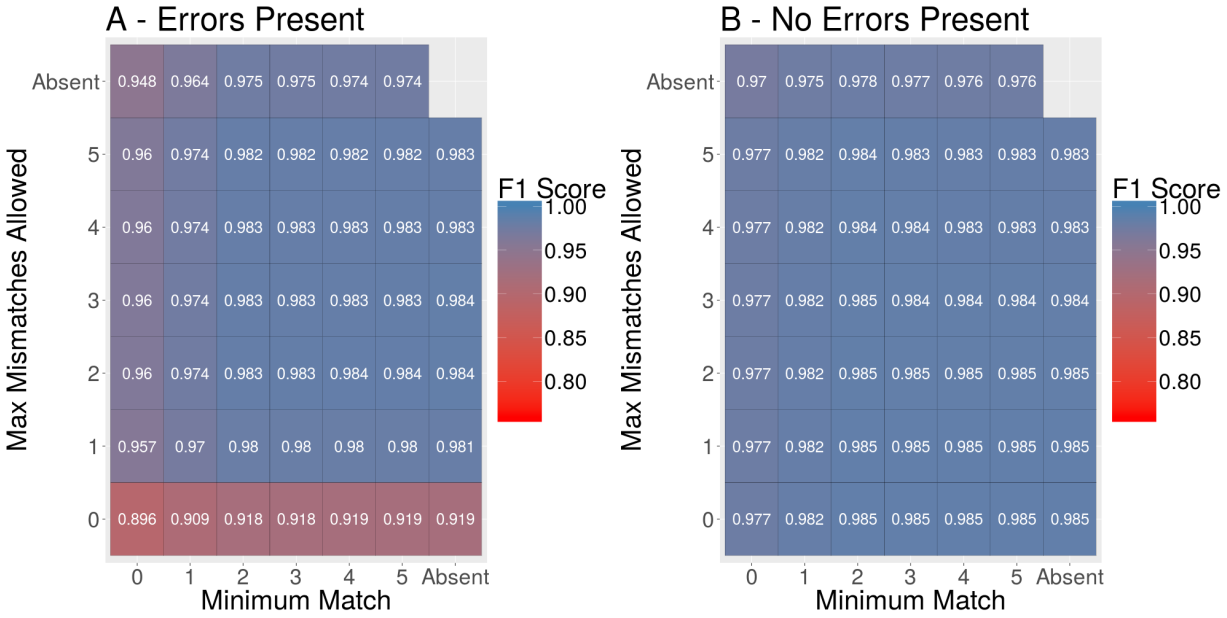
The key points of interest here are that the greatest gains to F1, and thus to the separability, are derived from allowing up to 2 multi-mappings, 2 mismatches and 3 imperfectly aligned bases. Whilst the combination of these scores does not result in the highest F1, Figure 2.3 shows that this combination of variables achieves an F1 of 0.971 where the highest achieved by any combination of variables is 0.98 which requires the same threshold for mismatch and imperfectly aligned bases, but without using the multi-map threshold. The performance of the separation for the 150bp was in keeping with these findings, as can be seen in the subsequent analysis in Figure 2.4, detailed further in the next paragraph. For the longer reads too the same relaxation of thresholds can be seen to achieve almost the highest F1 of any threshold combination, the biggest difference in F1 results between the read lengths can be seen to be the effect of a conservative mismatch threshold of 0 which can be seen to penalise the 150bp set more harshly simply as a result of its longer length and thus greater likelihood to contain an error.

Having assessed the effect of thresholding the key variables on Sargasso's species assignment, we moved forward to assess the effects of error presence and read length on assignment efficacy, in conjunction with mismatch and minimum match criteria. As the variable thresholds are less comparable for different lengths of reads, we have scaled each adjustment to the longer read set proportionately, with a threshold of one mismatch for 50bp now being up to 3 for 150bp. The results for this investigation can be seen in Figure 2.4 for the simulated rat reads aligned to the genomes of rat and mouse. For the simulated reads without error, the performance of the pipeline varies little with relaxation of the assignment criteria as the F1 scores are already very high. For the 50bp set we thus see that with a 2 step relaxation in the minimum match criteria we achieve the highest F1 score, notably only marginally (0.008) higher than the most conservative settings for the variables. For the 150bp set, we can lessen this to a 1 step relaxation and the difference between the maximum F1 is actually

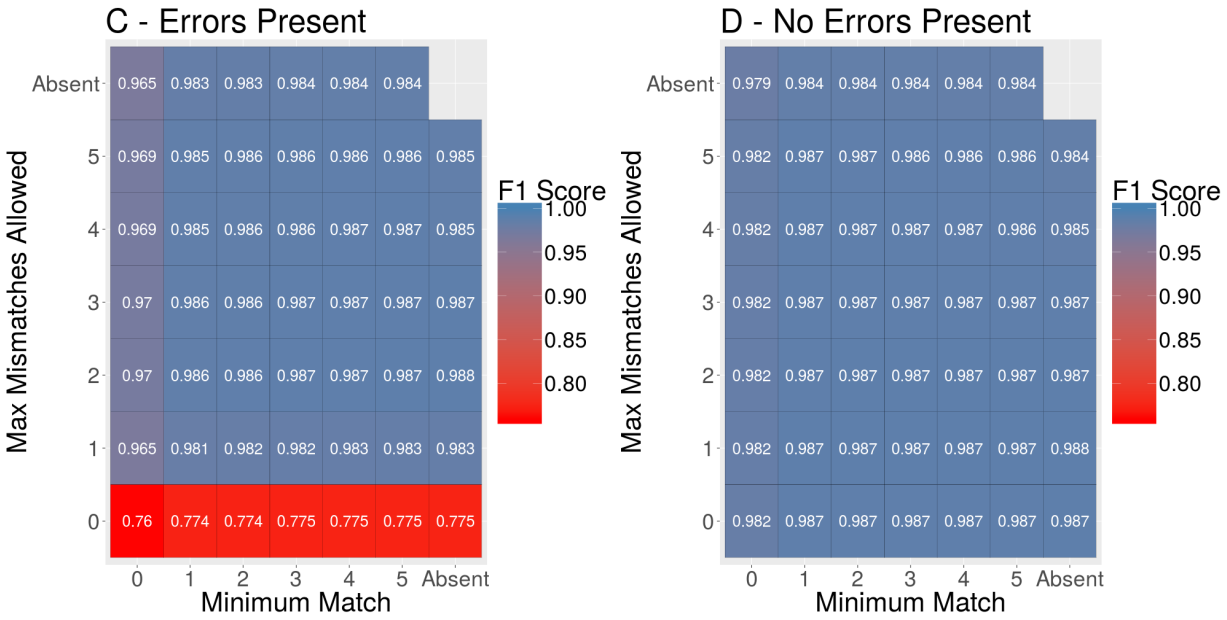
less (0.005). Mismatch relaxation for both read lengths only has a marginal negative impact on F1, though when it is absent and only minimum match is used we see a larger negative impact.

For the sets with sequence error however, there is a more obvious performance distinction between the 50bp and 150bp read length sets. Whilst relaxation of both key variables by one point accounts for the bulk of observed change in F1, the shorter read set is more sensitive with consistent change in F1 observable for a higher proportion of criteria variation. The longer read set performs notably worse for conservative variable values perhaps given that they are comparably more stringent for longer reads, subsequent threshold variations have been scaled proportionate to length. Longer reads will be more likely to contain instances of sequence inconsistency, compared to the reference genome, by simple virtue of their length providing more opportunity for such errors. Conversely it is their length that gives rise to a better performance with regard to F1 for more relaxed criteria, albeit marginal, and also for the lower quantity of reads classed as ambiguous as longer reads have less potential locations in both species for equal mapping. This decrease in ambiguity is to be expected as the longer a sequence's length, the less likely it is that there will be an identical stretch in another species' genome, even one genetically close. It is worth noting that the assignment made using only the mismatch criteria is almost equal in performance to assignment with decreasing the minimum match threshold by 5, thus showing that mismatch is perhaps the more important discriminator for this data.

These results demonstrate that Sargasso is particularly effective at distinguishing simulated RNA-Seq reads between two closely related species.



(a) **50bp** Reads



(b) **150bp** Reads

Figure 2.4: **Read Assignment for Simulated Rat Data for Variation in Filtering Criteria**. Minor relaxation in assignment thresholds is beneficial to both shorter and longer read lengths, though with a larger impact on the F1 for the 150bp reads (C), indeed relaxing the mismatch criteria by one step leads to a higher F1 for 150bp compared to the 50bp (A). Assignment thresholds for both read lengths in the absence of error (B,D) have little impact on assignment. The number of mismatches and minimum match bases is scaled proportionately for the 150bp data.

Species	Error				No Error			
	Stringent		Lenient		Stringent		Lenient	
	50bp	150bp	50bp	150bp	50bp	150bp	50bp	150bp
Mouse	0.903	0.762	0.983	0.986	0.983	0.987	0.988	0.991
Human	0.902	0.767	0.982	0.983	0.981	0.986	0.987	0.99
Zebrafish	0.889	0.757	0.977	0.971	0.97	0.976	0.984	0.99
σ	0.008	0.005	0.003	0.008	0.007	0.006	0.002	0.001

Table 2.2: **Performance of Sargasso Over Different Genetic Distances.** Sargasso can be seen to perform well on species of varying genetic distance, suffering only a minor decrease in F1 for particularly distant species. Simulated reads from the three species listed here was separated against the rat genome at both stringent filtering thresholds (no mismatches, full read alignment, no multimapping) and at more lenient thresholds (1 mismatch, up to 2 clipped bases, multimapping allowed) to highlight the largest point of variation in F1. Simulated data both with and without error was used. Standard deviation was calculated between the F1 scores of each species for variation in threshold and/or source data.

2.3.3 *Testing Simulated Data for Species of Varying Genetic Distance*

To check the Sargasso's consistency and capacity for generalisation we simulated three more sets of RNA-Seq data from different species that were then mapped to the genomes of both their own species (mouse, human and zebrafish) and to the genome of rat. Mouse was chosen as the pipeline's performance should be very similar to that of the Rat simulated set, human as it is a high quality mammal genome of further genetic distance and finally zebrafish as it is of great genetic distance. The results for variation in filter variable, read length and read error for these datasets so closely reflected the distribution of F1 for the rat separation in the previous section that I have summarised the F1 results at two key points of variation in Table 2.2.

Across all three species separations we can see a core consistency for F1 against both variation in filtering thresholds and in source data. This is demonstrated by a consistently low standard deviation across all points of variation tested. Whilst we can see that variation between the mammalian species is particularly low, there is a very minor decrease in performance for the zebrafish. However this could also be explained by comparative genome quality as this is lowest for the zebrafish of the three tested.

For the filtering thresholds we can see that the stringent thresholds produces a notably lower F1, particularly for the longer 150bp reads as simply by virtue of being longer they are more likely to include a sequencing error that prevents perfect align-

ment. For the lenient thresholds we can see a comparative performance to the reads without error.

Crucially what we observe here is that the effective operation of Sargasso is indeed consistent and generalisable across species of varying genetic distance, demonstrating comparable differentiating potential.

2.3.4 *Filtering Strategies*

Based on the results in the previous section, we made four particular choices of sets of filtering parameters available to users through a single command-line option: 'conservative', 'best', 'recall' and 'permissive'. In our 'conservative' strategy, precision is favoured over recall. Here, for a read or read pair to be assigned to a particular genome, there must be no mismatches with respect to the reference sequence, and the full length of the read must be mapped, with no clipping. Moreover, reads must map to a single locus in the reference genome. This strategy corresponds to the stringent thresholds in Section 2.3.3.

For our 'best' strategy, on the other hand, which provides a good balance between sensitivity and specificity, we note that the greatest gains in F1-score seen above (compared to the "conservative" strategy) are made when allowing at most 1 mismatch with respect to the reference genome for a 50bp paired-end read (or the same proportion of total read length for longer reads), and at most 2 base clipped for each single 50bp read in the pair (or the same proportion of each read's length for longer reads). This strategy corresponds to the lenient thresholds used in Section 2.3.3.

The 'recall' strategy has also been added in for users who wish to use strategy that prioritises recall over precision. This strategy thus prioritises the highest F1 scores, however will run the highest risk of false positives as a result of less stringent thresholds.

Lastly, the 'permissive' strategy maximises assignments to the species of origin at the expense of a high level of mis-assignments to the other species. This strategy has been added in to compare with the assignments of xenograft separation tools. A summary of the filtering strategies for easy comparison can be seen in Table 2.3.

2.3.5 *Computational Performance*

Figure 2.5 shows the time taken by Sargasso to process an RNA-seq data set comprising around 47.6 million 50bp paired-end reads. This time is inclusive of all of Sargasso's core stages of the pipeline: read mapping to each genome with `STAR`, sorting of mapped read files and assignment of reads to each genome. With a single

	Max No. Mismatches	Max. No Clipped Bases	Max No. Multi-maps	Overhang Threshold (bp)
Conservative	0	0	1	5
Best	1	2	999999	5
Recall	2	10	999999	5
Permissive	25	25	999999	0

Table 2.3: **Filtering Strategy Summary.** This table details the specific thresholds for each key assignment variable for each of Sargasso's four predefined filtering strategies.

processing core, reads are filtered by the pipeline at a rate of approximately 11.1 million reads per hour, with a total execution time of approximately 4.5 hours, this rises to around 86.6 million reads per hour processed with 24 cores, with a total execution time of approximately 35 minutes. The greatest speed improvements are made by increasing the number of cores to approximately 8, bringing the execution time to under an hour, with only marginal gains from further parallelisation made beyond this point. Beyond 20 cores, the difference is so marginal as to be of no significance.

2.3.6 *Simulation of Contamination Effects on Expression*

As the physical separation methods, to which we present Sargasso as an alternative, are known to contaminate separated data through the inclusion of non-desired cell-types, we simulated the result this would have downstream.

A cell-type contamination was achieved through addition of mRNA from a neuron monoculture sample into an astrocyte monoculture sample at a ratio of 95 : 5 prior to sequencing in order to achieve a 5% contamination rate. The abundance of gene expression, as FPKM, was then carried out on the sequenced RNA-Seq sample and compared with the result of an uncontaminated astrocyte monoculture sample cultured under the same conditions, this can be seen in Figure 2.6.

As we can see, the contamination of a monoculture dataset by as little as 5% of cells from another type has a significant effect on downstream analysis. In Figure 2.6 we can see total of 863 genes are expressed more than two-fold higher, for which 216 of these genes are expressed more than ten-fold higher, as a result of this contamination.

2.3.7 *Impact of Sargasso on Downstream Analysis*

In this section I will discuss the results of our investigations into the impact of Sargasso's usage on the results of downstream analysis. These investigations will cover read loss, incorrect assignment, changes to differential gene expression and the detection of key experimental genes.

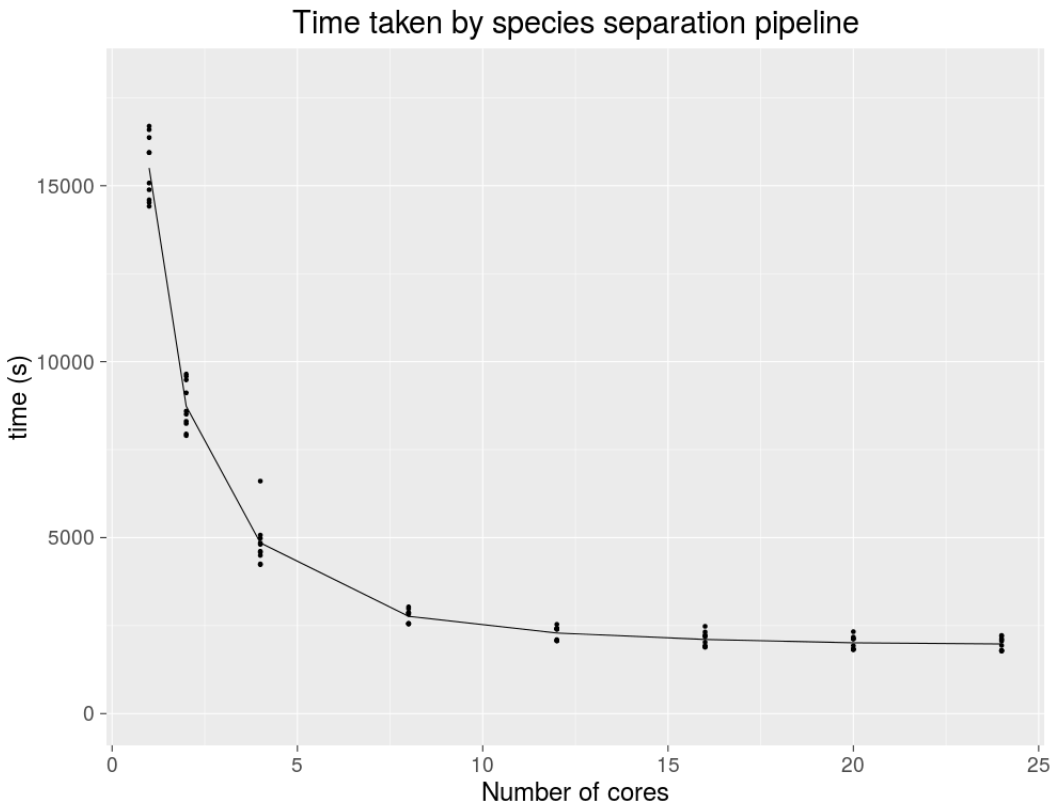


Figure 2.5: **Speed of Sargasso Execution.** Sargasso's total execution time increases greatly with parallelisation up until 8 threads are used, after which point execution takes less than an hour and gains are marginal. For this analysis an RNA-seq data set comprising 47.6 million 50bp paired-end reads was passed through the species separation pipeline using 1, 2, 4, 8, 12, 16, 20 or 24 threads, and the wall clock time used was recorded. The mean time taken over 10 repetitions is shown. Figure taken from Heron, Dando & Simpson, *forthcoming*.

2.3.7.1 *Read Loss*

It is important to note that the loss of reads due to ambiguity of their species of origin will not be equally distributed across the genome and this could affect downstream analysis. For example, genes which are highly conserved between two species may be subject to greater loss of reads than those which are not conserved, and this might impact discovery of differential expression in such genes.

To investigate these potential effects further, we first generated all possible theoretical paired-end reads of length 50 base pairs, using an insert size of 150 base pairs, from mouse and rat genes marked as protein coding in Ensembl version 84. We then passed these reads through Sargasso using both the 'conservative' and 'best' filtering strategies described in the previous section.

Figure 2.7(a) shows the per-gene fractions of all theoretical reads from the protein coding mouse transcriptome lost due to species separation, when aligned to both mouse and rat by Sargasso. As can be seen, for the great majority of genes, the

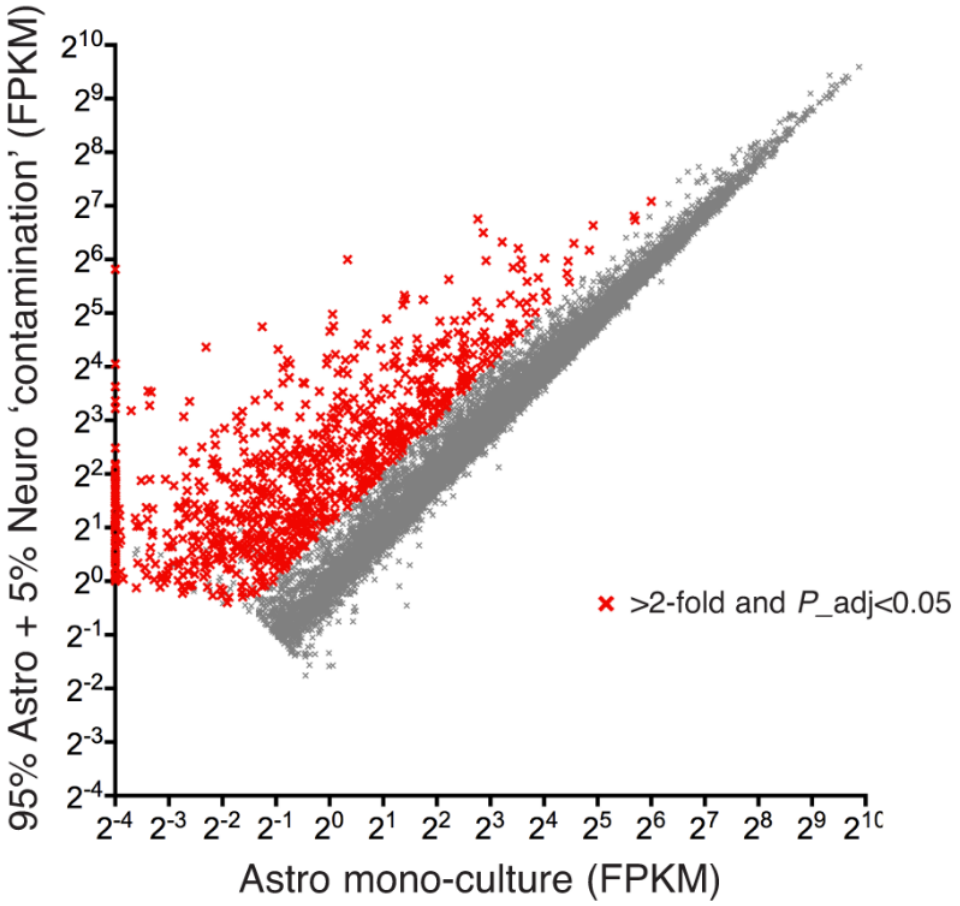


Figure 2.6: **Change in FPKM of Astroycte Genes as a Result of 5% Neuronal Contamination**. A cell-type contamination of just 5% results in the expression of 863 genes to be increased by more than two-fold. This figure shows the change in abundance of gene expression as a result of cell-type contamination, in this case an astrocyte sample contaminated at a ratio of 95 : 5 with neuronal RNA. The FPKM of this contaminated sample can be seen here plotted against the FPKM of an uncontaminated astrocyte sample. Figure reproduced from (Hasel et al., 2017): Supplementary Figure 3A.

majority of reads can be unambiguously assigned to mouse: for the "conservative" strategy 85.7% of genes lose fewer than 20% of reads, with 91.6% of all theoretical reads unambiguously assigned to the mouse genome; for the "best" strategy 98.9% of genes lose fewer than 20% of reads, with 96.3% of total reads assigned to the mouse. Similarly, 2.7(b) shows the per-gene fractions of all theoretical reads from the protein coding rat transcriptome lost due to species separation, when aligned to both rat and mouse. Here, for the "conservative" strategy, 79.4% of genes lose fewer than 20% of reads, with 86.5% of all theoretical reads unambiguously assigned to the rat, and for the "best" strategy this rises to 98.4% (with 95% of total reads unambiguously assigned). The gap in performance here is likely not due to Sargasso's methodology but to the comparative quality of the two species' reference genomes, with the mouse

genome being higher as a result of it being a more frequent subject of study. For the rat genome there are more sequence gaps and less annotation meaning that when the mouse reads are aligned to both genomes, there will be more instances where there is not a comparable location yet annotated in the rat genome thus allowing for unambiguous assignment and, overall, better results for the mouse than the rat reads.

When analysing differential gene expression reads are assigned to gene features and counted (at this point it is standard practice to discard reads that multi-map to the genome). Hence we used the read summary tool `featureCounts` in order to count the number of theoretical reads assigned to genes after Sargasso has separated the data by species, for both the 'conservative' and 'best' strategies. We then compared the per-gene counts with those obtained after performing a standard read mapping to a single genome with `STAR`.

Figure 2.8(a) shows the per-gene fractions of feature-assigned theoretical reads from the protein coding mouse transcriptome that are lost due to species separation with Sargasso, when mapped to mouse and rat, when compared to counts of reads assigned after a standard `STAR` mapping. In this case, for the 'conservative' strategy, only 5.2% of genes lose more than 20% of reads when compared to standard `STAR` mapping, and for the 'best' strategy, this drops to 0.4%. Similar behaviour for the rat, when mapped to rat and mouse, can be seen in Figure 2.8(b). Here 5.4% of genes lose more than 20% of reads in the 'conservative' strategy, and 1.4% for the 'best' strategy. Thus we can see that only a small amount of data is lost for the great majority of genes as a result of Sargasso's application.

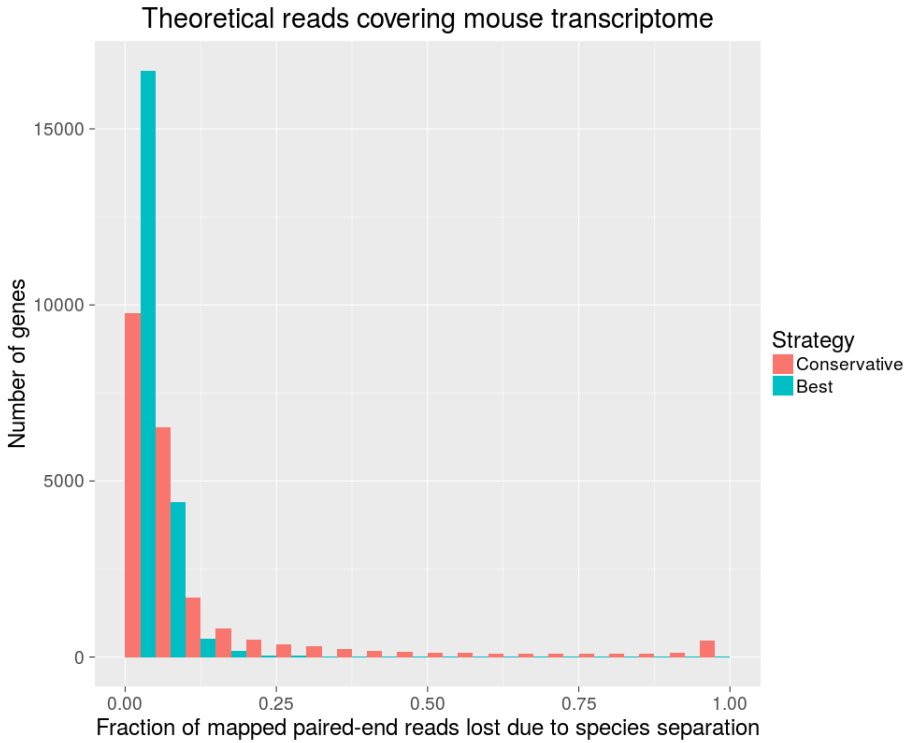
We were interested in investigating whether the genes which lose a larger fraction of their reads as a result of Sargasso's application share any particular features. We thus used the R package topGO to perform a gene ontology analysis in order to look for enriched biological processes in those genes which lose more than 20% of their feature-assigned reads when compared to standard `STAR` mapping. The topGO analysis used the 'elim' methodology and for a background set it used all protein-coding genes in Ensembl version 84, all genes which lost more than 20% of their feature-assigned reads as a result of Sargasso were tested against this. Multiple testing was not carried out on the resultant p-values in accordance with the authors' justification: p-value calculation for each term is conditioned on neighbouring terms and as a result are not independent, thus multiple testing theory does not directly apply (Alexa and Rahnenführer, 2018).

For the mouse transcriptome with the 'conservative' filtering strategy, for which 1153 genes lose more than 20% of reads, the significantly enriched ($p<0.05$) GO term of greatest interest, with joint lowest p-value and the second highest number of significant genes, was "G-protein coupled receptor signaling pathway" (GO:0007186,

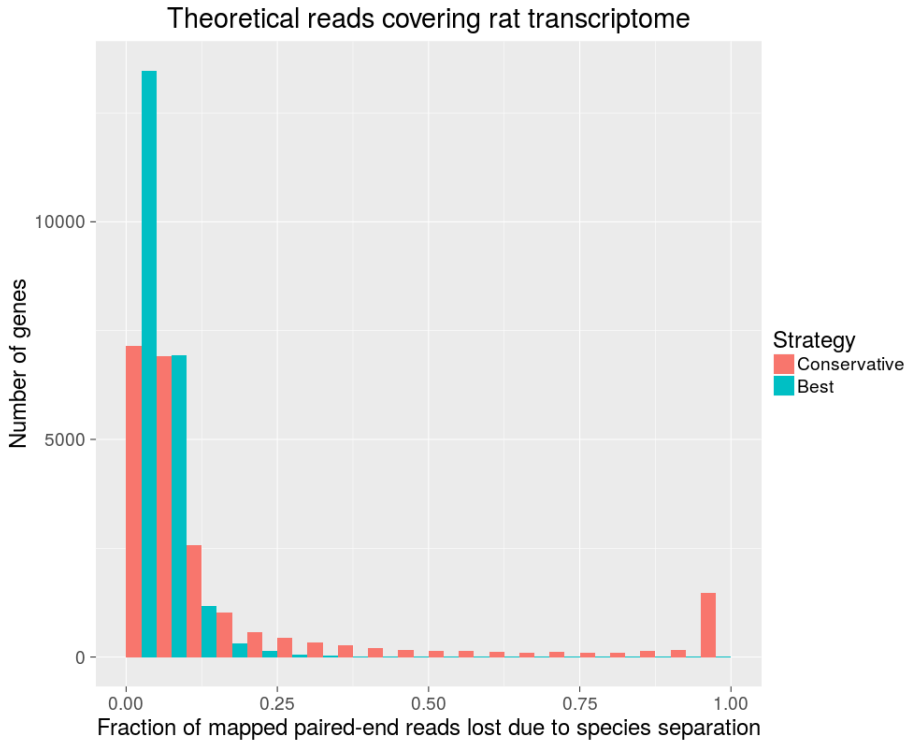
290/93.76 genes (significant/expected), p<1e-30). Of these 201 (66%) are marked in Ensembl version 84 as having more than one rat ortholog (the true number likely being higher, due to incomplete annotation of the rat genome) - indeed 53.9% (622 out of 1153) of mouse genes which lose more than 20% of reads are marked as having more than one rat ortholog. With our 'conservative' strategy genes whose reads have only a single mapping to the mouse genome, but who have multiple mappings to the rat genome, are those most likely to lose a large proportion of their reads.

However, this is not the case for the 'best' filtering strategy where only 88 genes lose more than 20% of their feature-assigned reads; of these only 4 are marked as having more than one rat ortholog. Here, instead, 73.9% of mouse genes (65 out of 88) are marked as having single rat ortholog. The GO term with the greatest number of significant genes (p <0.05) is "regulation of transcription, DNA-templated" (GO:0006355, 26/10.75 genes (significant/expected), p=0.00966). The genes that lose a large proportions of reads when using this strategy tend to be those with a high level of sequence conservation between species.

The topGO enrichment results for both filtering strategies can be found in Appendix B in Tables B.1 & B.2.

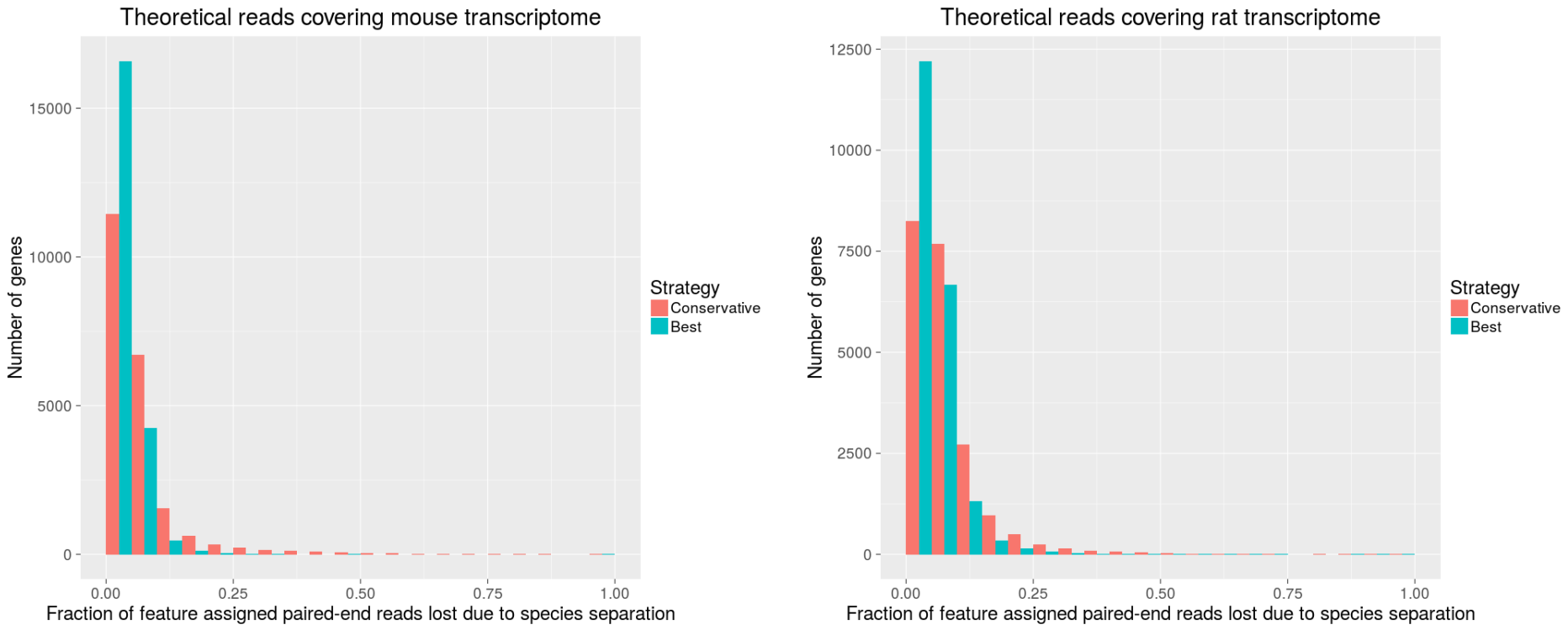


(a) Mouse Reads



(b) Rat Reads

Figure 2.7: **Mapped reads lost due to species separation for Mouse and Rat.** Only a minority of genes lose more than 25% of reads mapped from both species of origin, mouse (a) or rat (b), as a result of Sargasso for the stringent 'conservative' filtering strategy. Use of Sargasso's 'best' strategy can be seen to minimise this loss. Here all possible 50 base pair paired-end reads (with insert size 150 base pairs) from the entire protein-coding transcriptome in Ensembl version 84, for both mouse (mapped against mouse and rat) and rat (mapped against rat and mouse) were generated and separated by Sargasso with 'conservative' and 'best' strategies. The fraction of reads lost due to species separation was calculated on a per-gene basis and a frequency distribution histogram generated (5% bins). Figures taken from Heron, Dando & Simpson, *forthcoming*.



(a) Mouse Reads

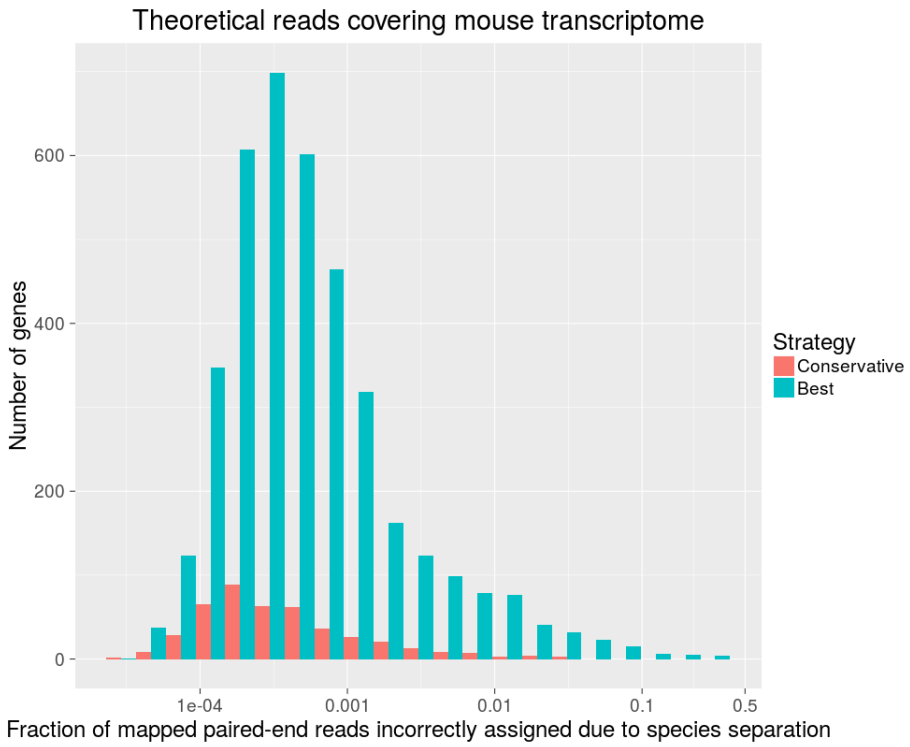
(b) Rat Reads

Figure 2.8: **Feature-assigned reads lost due to species separation**. For the 'conservative' filtering strategies, a minority of features can be seen to lose more than 15% of their reads for both mouse (a) and rat (b) simulated reads. This again can be decreased through the use of the 'best' strategy. Here all possible 50 base pair paired-end reads (with insert size 150 base pairs) from the entire protein-coding transcriptome in Ensembl version 84, for both mouse (mapped against mouse and rat) and rat (mapped against rat and mouse) were generated and separated by Sargasso with 'conservative' and 'best' strategies. Per-gene read counts were calculated using the `featureCounts` tool. The fraction of reads lost due to species separation, when compared with read counts obtained after standard `STAR` mapping, was calculated on a per-gene basis and a frequency distribution histogram generated (5% bins). Figures taken from Heron, Dando & Simpson, *forthcoming*.

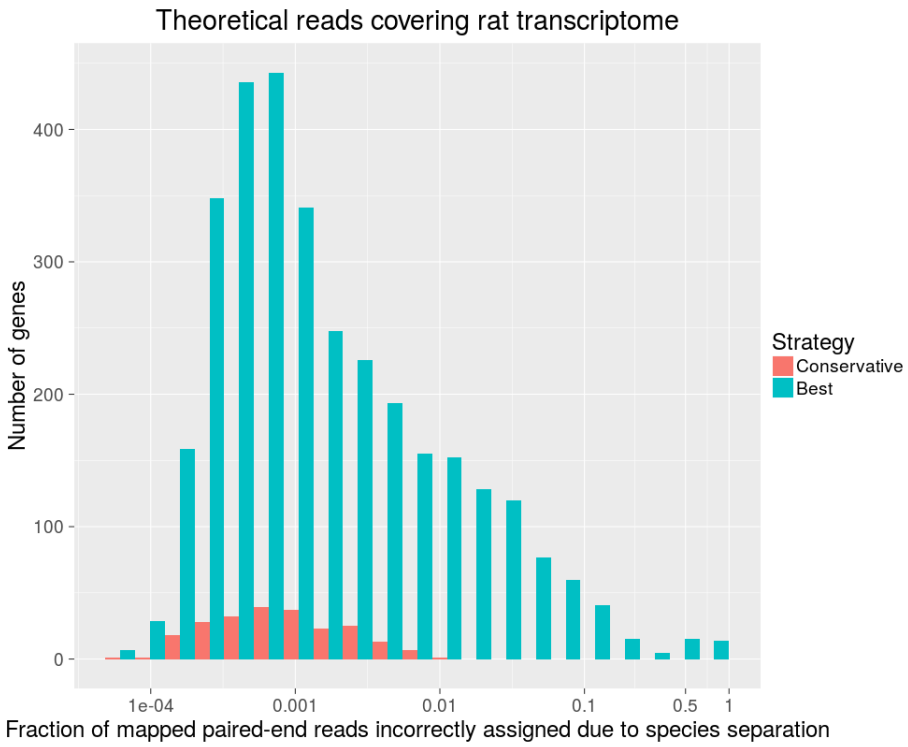
2.3.7.2 *Incorrect Assignment*

In addition to reads which the species separation pipeline discards due to ambiguity of their species of origin, it is also important to consider reads that might be incorrectly assigned to the wrong species, as these will bias downstream analysis. Figure 2.9(a) shows the per-gene fractions of all theoretical reads from the protein coding mouse transcriptome incorrectly assigned to the rat genome by Sargasso when mapping against mouse and rat. With our 'conservative' strategy, which prioritises minimising the number of reads mis-assigned to the wrong species, only 10 genes (0.05%) have more than 1% of theoretical paired-end reads incorrectly assigned, with 0.002% of all theoretical mouse reads wrongly allocated to the rat genome. With our 'best' strategy, which balances precision and recall, this rises to 195 genes (0.9%) with more than 1% of reads incorrectly assigned, with 0.03% of total reads mis-allocated.

In the case of the protein coding rat transcriptome in Figure 2.9(b), when mapping against rat and mouse, only 1 gene has more than 1% of its reads mis-assigned for the 'conservative' strategy, with 0.001% of all theoretical rat reads wrongly allocated to the mouse genome. However, with the 'best' strategy this rises to 585 genes (2.6%), with 38 genes (0.2%) having more than 20% of their reads incorrectly assigned and 0.1% of total rat reads mis-assigned. Of those 585 genes with more than 1% of reads mis-assigned, 291 (49.7%) are marked in Ensembl version 84 as being orthologs of single mouse genes which themselves have more than one rat ortholog. Whether in such cases the multiple rat genes represent true paralogs, or perhaps mis-assemblies in the rat genome, any reads from these genes which map equally well to both genomes will be incorrectly assigned to the mouse due to having fewer multi-mapping loci.



(a) Mouse Reads



(b) Rat Reads

Figure 2.9: **Reads incorrectly assigned to the mouse and rat genomes**. Very few reads are assigned incorrectly for reads belonging to either mouse (a) or rat (b), this number can be minimised greatly through use of the 'conservative' strategy. Incorrect assignment remains however less than 0.01% for the majority of genes affected even when using the 'best' strategy. From the entire protein-coding transcriptome in Ensembl version 84, for both mouse (mapped to mouse and rat) and rat (mapped to rat and mouse), all possible 50 base pair paired-end reads (with insert size 150 base pairs) were generated and passed through Sargasso with 'conservative' and 'best' strategies. The fraction of reads incorrectly assigned to the mouse genome was calculated on a per-gene basis and a frequency distribution histogram generated (5% bins). Figures taken from Heron, Dando & Simpson, *forthcoming*.

2.3.7.3 *Impact on Differentially Expressed Genes*

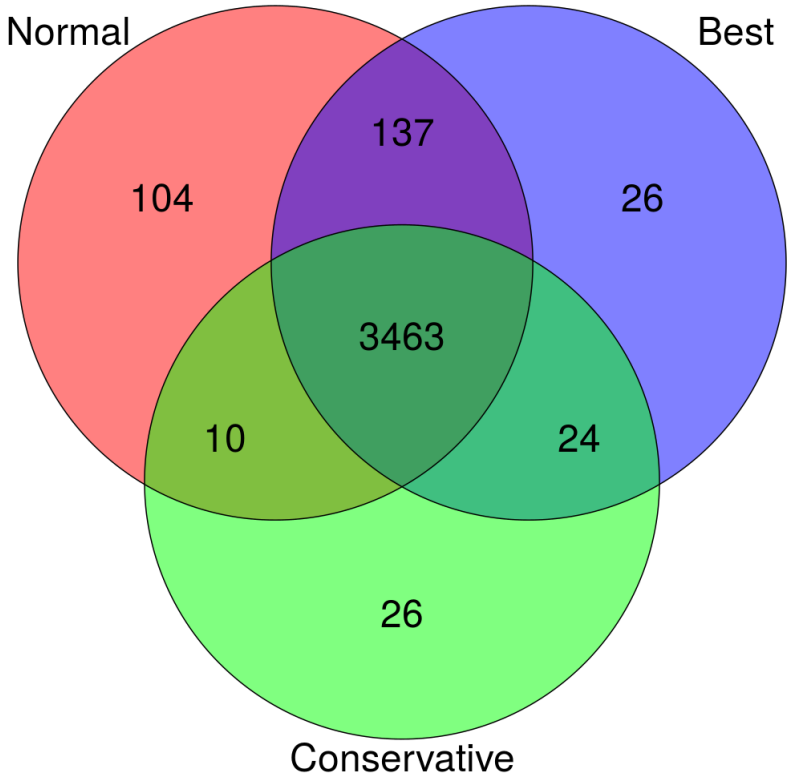
Finally, we went on to consider the effect of Sargasso's filtering on downstream differential gene expression analysis itself. For this investigation we used a single species (mouse) RNA-Seq dataset, ADm-MN, sequenced from an *in vitro* neuron monoculture before and after treatment with bicuculline+4-aminopyridine (n=3 in each condition) to induce synaptic activity and concomitant gene expression changes (Hasel et al., 2017). This data was passed through Sargasso, mapping to mouse and rat using both 'conservative' and 'best' filtering strategies, and mapped reads were assigned to genes and counted using `featureCounts`. We then performed differential gene expression using DESeq2, and compared results after using our two filtering strategies with those obtained from differential expression analysis after a normal mapping with `STAR`.

Figures 2.10(a) and 2.10(b) show the overlaps of genes called as up- and down-regulated (FDR<0.1) in the three cases. For the up-regulated genes, of the 3714 called as differentially expressed after using normal `STAR` mapping, 114 genes (3.1%) are no longer called differentially expressed when using our 'best' filtering strategy, while 50 genes are newly called as differentially expressed. For the 'conservative' filtering strategy this rises to 241 genes (6.5%) no longer called differentially expressed, with, again, 50 genes newly called as up-regulated. For the down-regulated genes, of the 3376 called as differentially expressed after normal `STAR` mapping, 110 genes (3.3%) are no longer differentially expressed for 'best', whilst 34 new genes being called as differentially down-regulated. For the 'conservative' filtering strategy this rises to 306 genes (9.1%) no longer differentially expressed with 27 genes newly called as differentially down-regulated.

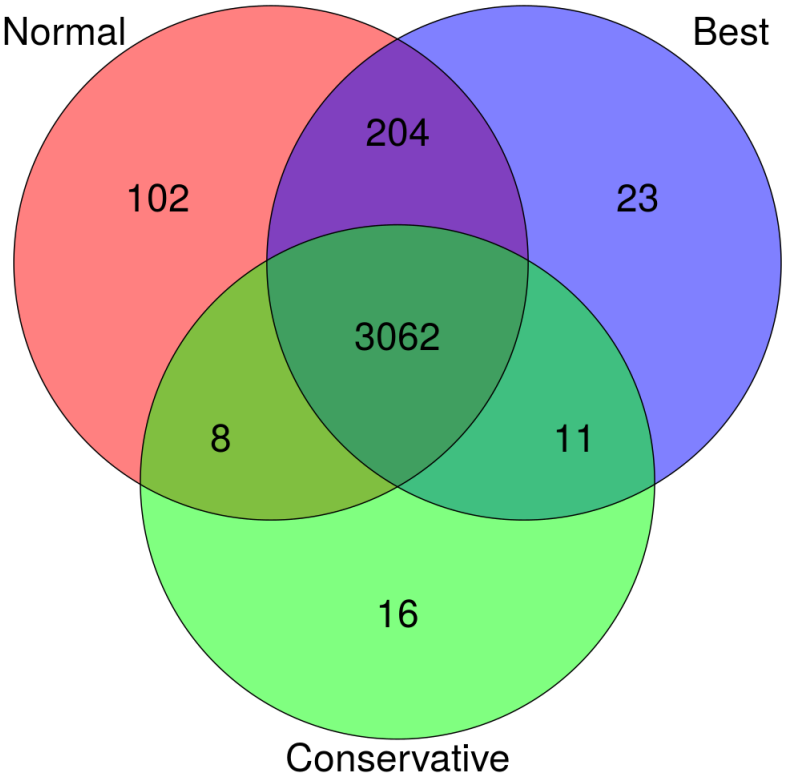
The changes here are reassuringly small for the 'best' filtering strategy, however with almost 10% of differentially expressed down-regulated genes missing for the 'conservative' strategy, this loss of sensitivity will have to be factored in when making parameter choices.

Figures 2.11(a) and 2.11(b) plot the gene fold changes due to induction of synaptic activity, calculated from RNA-seq data subject to our 'best' and 'conservative' filtering strategies, when mapped to both mouse and rat genomes, as compared to those derived by DESeq2 after normal `STAR` mapping, focussing on those 12304 genes with >1 mean FPKM after normal `STAR` mapping. As can be seen, passing data through our species separation pipeline has little effect on the fold changes calculated, with correlations of 0.999 and 0.998 respectively when comparing the 'best' and 'conservative' strategy fold changes with those after normal `STAR` mapping.

Thus whilst we do observe some loss in the number of genes being called as differentially expressed, we can be confident that detection of key transcriptomic changes, instigated experimentally, are barely affected by the use of Sargasso.

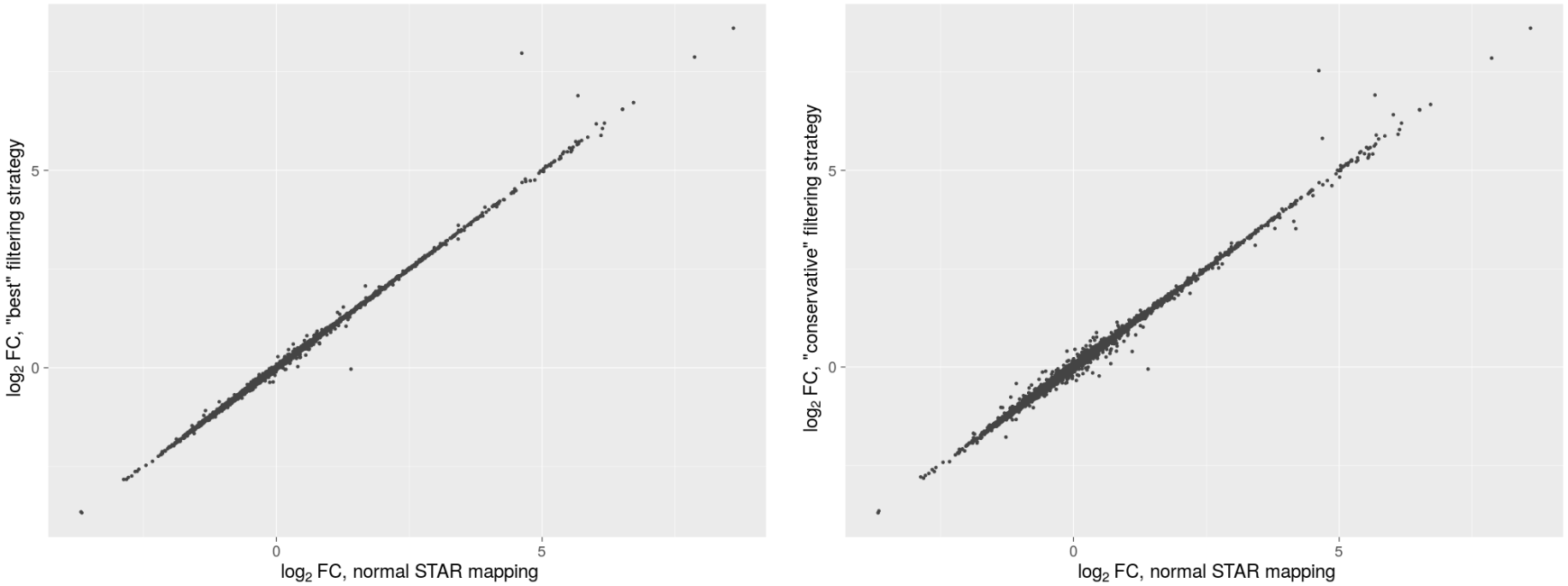


(a) Up-Regulated Differentially Expressed Genes



(b) Down-Regulated Differentially Expressed Genes

Figure 2.10: **Overlap of up- and down-regulated differentially expressed genes**. Sargasso's 'best' strategy can be observed to have a minimal effect on downstream differential expression analysis whilst use of the 'conservative' strategy results in a small drop in reporting due to its more stringent thresholds. RNA-seq was performed on ADm-MN, *in vitro* monocultured mouse neurons before and after the induction of synaptic activity, and differentially expressed genes calculated in three cases: (i) normal RNA-seq processing pipeline, (ii) with Sargasso (mapping to mouse and rat genomes) using our 'best' filtering strategy, and (iii) with Sargasso (mapping to mouse and rat genomes) using our 'conservative' strategy. The Venn diagram in (a) shows the overlap of genes called as up-regulated (FDR<0.1) and (b) shows the overlap of genes called as down-regulated (FDR<0.1). Figures taken from Heron, Dando & Simpson, *forthcoming*.



(a) 'best'

(b) 'conservative'

Figure 2.11: **Altered gene fold changes with 'best' and 'conservative' filtering strategies**. Sargasso's 'best' strategy (a) can be observed to have a minimal effect on fold change whilst the use of the 'conservative' strategy (b) has a greater effect both positively and negatively affecting the fold change of a minority of genes due to its more stringent thresholds. RNA-seq was performed on ADm-MN, mouse neurons before and after the induction of synaptic activity, and differentially expressed genes calculated. The fold-changes derived from our 'best' and 'conservative' filtering strategies, mapping against mouse and rat genomes, were plotted against those after normal `STAR` mapping, focussing on the 12304 genes with > 1 mean FPKM after normal `STAR` mapping. Figure taken from Heron, Dando & Simpson, *forthcoming*.

2.3.8 *Impact of Sargasso on Protein Coding Reads*

As protein coding genes are well conserved between species, we decided to test the performance of Sargasso specifically on reads from only protein coding genes to see whether assignment would be different for this subset of reads. For this analysis, we mapped reads from a full monoculture dataset of mouse astrocytes, OSm-MA, to the genomes of both rat and mouse with Sargasso using both 'conservative' and 'best' filtering strategies. This dataset is detailed further in Section 1.4.1. The assignment results can be seen in Table 2.4.

The reads from the initial `STAR` mapping and the filtered reads output by Sargasso were then quantified with `featureCounts` and the counts for protein coding reads were totalled. These totals were then contrasted to derive the number of protein coding reads mapped to the genome of each species that did not make it through Sargasso's filtering process. The results can be seen in Table 2.5.

We can see that the proportion of coding reads lost for both strategies is in fact lower than that for all reads in Table 2.4 and is consist across all conditions and replicates. However the use of only uniquely mapping reads here could be a factor to the lower proportion of loss. These results demonstrate that Sargasso's performance is consistent when restricted to only reads from protein coding genes.

OSm-MA 'conservative'		Mouse				Rat			
Condition	Replicate	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous
Control	1	60977266	41098911 (**67.4%**)	19170143 (**31.4%**)	708212 (**1.2%**)	27895023	33478 (**0.1%**)	27153333 (**97.3%**)	708212 (**2.5%**)
Control	2	52500890	35345444 (**67.3%**)	16524745 (**31.5%**)	630701 (**1.2%**)	24419535	32173 (**0.1%**)	23756661 (**97.3%**)	630701 (**2.6%**)
Control	3	65881534	44588385 (**67.7%**)	20500236 (**31.1%**)	792913 (**1.2%**)	30242026	45635 (**0.2%**)	29403478 (**97.2%**)	792913 (**2.6%**)
4h Post Stimulus	1	48800341	32519505 (**66.6%**)	15708015 (**32.2%**)	572821 (**1.2%**)	23226394	30415 (**0.1%**)	22623158 (**97.4%**)	572821 (**2.5%**)
4h Post Stimulus	2	44788426	30376221 (**67.8%**)	13850168 (**31.0%**)	562037 (**1.3%**)	20975373	28376 (**0.1%**)	20384960 (**97.2%**)	562037 (**2.7%**)
4h Post Stimulus	3	46278125	31397930 (**67.9%**)	14310078 (**30.9%**)	570117 (**1.2%**)	21765856	30460 (**0.1%**)	21165279 (**97.2%**)	570117 (**2.6%**)
24h Post Stimulus	1	65354376	44328564 (**67.8%**)	20238516 (**31.0%**)	787296 (**1.2%**)	30225776	46631 (**0.2%**)	29391849 (**97.2%**)	787296 (**2.6%**)
24h Post Stimulus	2	62820193	42341896 (**67.4%**)	19705370 (**31.4%**)	772927 (**1.2%**)	29310771	47406 (**0.2%**)	28490438 (**97.2%**)	772927 (**2.6%**)
24h Post Stimulus	3	58195816	39165436 (**67.3%**)	18339389 (**31.5%**)	690991 (**1.2%**)	27033933	41199 (**0.2%**)	26301743 (**97.3%**)	690991 (**2.6%**)
OSm-MA 'best'		**Mouse**				**Rat**			
Condition	Replicate	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous
Control	1	60977266	56724544 (**93.0%**)	3337814 (**5.5%**)	914908 (**1.5%**)	27895023	522968 (**1.9%**)	26457147 (**94.9%**)	914908 (**3.3%**)
Control	2	52500890	48769320 (**92.9%**)	2913852 (**5.6%**)	817718 (**1.6%**)	24419535	572476 (**2.3%**)	23029341 (**94.3%**)	817718 (**3.4%**)
Control	3	65881534	61267081 (**93.0%**)	3593124 (**5.5%**)	1021329 (**1.6%**)	30242026	678226 (**2.2%**)	28542471 (**94.4%**)	1021329 (**3.4%**)
4h Post Stimulus	1	48800341	45252094 (**92.7%**)	2790855 (**5.7%**)	757392 (**1.6%**)	23226394	623459 (**2.7%**)	21845543 (**94.1%**)	757392 (**3.3%**)
4h Post Stimulus	2	44788426	41503670 (**92.7%**)	2561294 (**5.7%**)	723462 (**1.6%**)	20975373	473321 (**2.3%**)	19778590 (**94.3%**)	723462 (**3.5%**)
4h Post Stimulus	3	46278125	42884026 (**92.7%**)	2658784 (**5.8%**)	735315 (**1.6%**)	21765856	527832 (**2.4%**)	20502709 (**94.2%**)	735315 (**3.4%**)
24h Post Stimulus	1	65354376	60607768 (**92.7%**)	3727930 (**5.7%**)	1018678 (**1.6%**)	30225776	824109 (**2.7%**)	28382989 (**93.9%**)	1018678 (**3.4%**)
24h Post Stimulus	2	62820193	58162628 (**92.6%**)	3655241 (**5.8%**)	1002324 (**1.6%**)	29310771	768438 (**2.6%**)	27540009 (**94.0%**)	1002324 (**3.4%**)
24h Post Stimulus	3	58195816	53869783 (**92.6%**)	3426952 (**5.9%**)	899081 (**1.6%**)	27033933	701319 (**2.6%**)	25433533 (**94.1%**)	899081 (**3.3%**)

Table 2.4: **Sargasso Read Assignment for OSm-MA.** In this table, Sargasso can be seen to effectively assign reads to their species of origin from the monoculture OSm-MA dataset, for both 'conservative' and 'best' filtering approaches. The filtering strategies perform as intended with the 'conservative' assigning less reads but with only a severe minority, an average of less than 0.2%, of incorrect assignments to the rat and the 'best' correctly assigning almost the entirity of the dataset to the mouse but with a higher level of incorrect assignment, an average of 2.4%. OSm-MA is comprised of 3 conditions with 3 replicates. The results from this assignment were used to assess protein coding reads lost as a result of Sargasso's application.

OSm-MA		Mapped to Genome		Sargasso 'conservative'		Percentage Lost		Sargasso 'best'		Percentage Lost	
Condition	Replicate	Mouse	Rat	Mouse	Rat	Mouse	Rat	Mouse	Rat	Mouse	Rat
Control	1	51440812	21024900	37054179	13145	28.0	99.9	48064701	53489	6.6	99.8
Control	2	44103765	18259986	31836880	11903	27.8	99.9	41213639	48586	6.6	99.7
Control	3	55092354	22514024	40011086	19149	27.4	99.9	51522450	64047	6.5	99.7
4h Post Stimulus	1	41197444	17368133	29323464	9856	28.8	99.9	38478235	41667	6.6	99.8
4h Post Stimulus	2	37643912	15868717	27221491	8694	27.7	99.9	35094001	38087	6.8	99.8
4h Post Stimulus	3	39005747	16546012	28206330	8754	27.7	99.9	36367562	36874	6.8	99.8
24h Post Stimulus	1	54499174	22528266	39528605	16387	27.5	99.9	50897172	59446	6.6	99.7
24h Post Stimulus	2	52472959	21890357	37835080	17021	27.9	99.9	48920559	61462	6.8	99.7
24h Post Stimulus	3	48714215	20323612	35029119	13050	28.1	99.9	45389948	49089	6.8	99.8

Table 2.5: **Sargasso Assignment of Protein Coding Reads for OSm-MA.** Sargasso's impact on protein coding reads can be seen to reflect the overall proportion of reads lost through Sargasso's application as previously demonstrated both in this section and more widely in this chapter. This was assessed through the application of Sargasso to a full dataset of monoculture astrocyte samples: OSm-MA, comprising of 3 conditions with 3 replicates. Both Sargasso's 'conservative' and 'best' filtering strategies were applied. The figures here describe uniquely mapping reads only, with reads lost here including both rejected and ambiguous reads.

2.3.9 *Evaluation with Cultured Stem Cell Data*

Having demonstrated the performance of Sargasso on simulated RNA-Seq data, we compared this performance to that for samples from the AD dataset. The first set is a 50bp control of solely rat neuron cells (ADm-RN) and the second is from a co-culture where each cell-type is grown from stem cells of a different species: rat (ADcc-RN) and mouse (ADcc-MA). As the simulated datasets generated by Flux Simulator, used in Sections 2.3.1 and 2.3.3, are small compared to the experimental datasets, I have provided a subsample of the ADm-RN sample which I have used as a more direct comparison with the simulated data in this section.

This data thus presents an opportunity to evaluate Sargasso's performance on real *in vitro* experimental data, with the monoculture serving as a control both in the context of the experimentation and for Sargasso's assignment. The results of Sargasso's species assignment can be seen summarised in Table 2.6 for both ADm-RN, full-size and sub-sampled, and ADcc for mapping against mouse and rat genomes. The F1 scores for variance of the key separation variables can be seen for the sub-sampled ADm-RN in contrast to the simulated rat reads in Figure 2.12. The simulated reads are the same dataset used previously and so the results for this data are the same as in Figure 2.4 and are presented again here for ease of comparison.

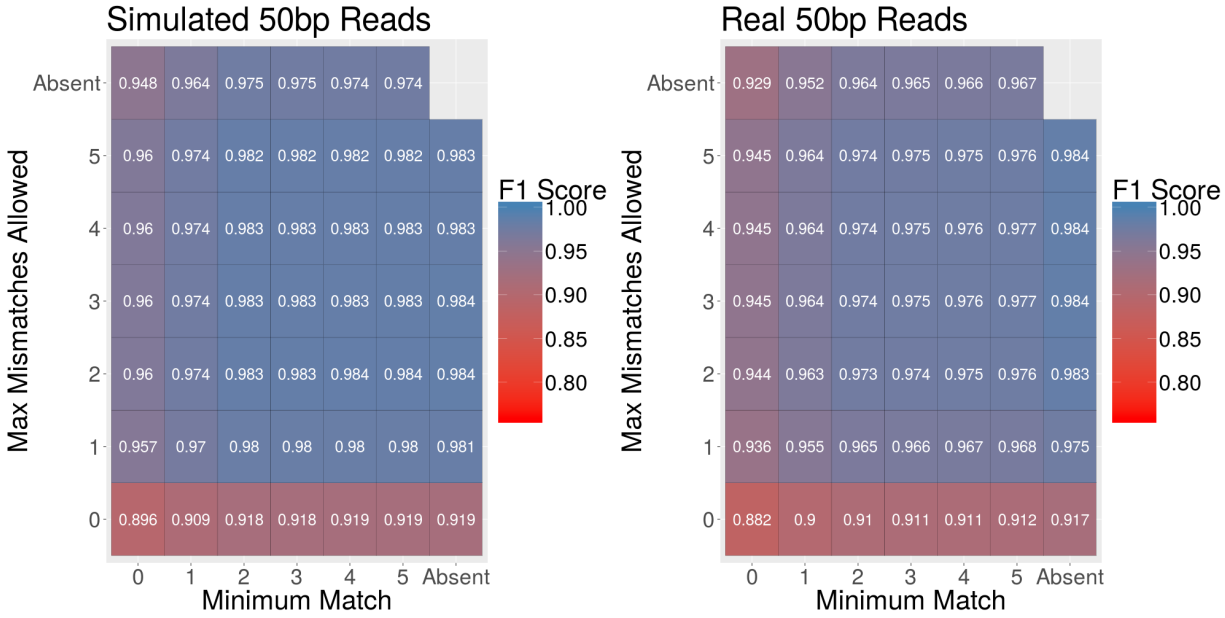


Figure 2.12: **Read Assignment for Simulated RNA-Seq Compared to Experimental Dataset, RatOnly**. Sargasso's performance is comparable for both simulated and real RNA-Seq reads. This figure shows the change in F1 Score for variation in the key separation variables, mismatches and minimum match, for experimental RNA-Seq data from ADm-RN (Sub-Sample) in contrast to simulated RNA-Seq data for 50bp rat reads. Both datasets were mapped by Sargasso to both the mouse and rat genomes.

	Rat				Mouse					
	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous	Precision	Recall
ADm-RN (Sub-Sample)	460550	356574 (**77.4%**)	95167 (**20.7%**)	8809 (**1.9%**)	212859	342 (**0.2%**)	203708 (**95.7%**)	8809 (**4.1%**)	0.9990418	0.7893328
ADm-RN	45986002	35620014 (77.5%)	9473877 (20.6%)	892111 (1.9%)	21231530	33929 (0.16%)	20305490 (95.6%)	892111 (4.2%)	0.9990484	0.7899078
ADcc	141633485	106691105 **(75.3%)**	32189451 **(22.7%)**	2752929 **(1.9%)**	105999393	47333406 **(44.7%)**	55913058 **(52.7%)**	2752929 **(2.6%)**	N/A	N/A

Table 2.6: **Read Assignment for Mono and Co-culture Experimental Data Using Sargasso.** Read assignment on experimental monoculture data closely replicates the level of separation observed for simulated data. Separation of a co-culture sample shows an effective, though uneven, partioning of the data that likely reflects relative cell-type population in culture. This table details the quantity of reads assigned to, rejected from and unassigned to the mouse or rat genome for ADm-RN and replicate 1 of the control condition of ADcc. ADm-RN (Sub-Sample) is a sampled down version of the subsequently listed ADm-RN, to approximately 1% of the size of the latter, this is displayed here to show comparative performance. The ADm-RN dataset consists of reads of 50bp length sequenced from monocultured rat neurons. The ADcc dataset consists of reads of 50bp length sequenced from a mixed-species co-culture of rat neurons and mouse astrocytes. The table lists N/A for the precision and recall of the co-culture as we cannot be certain, due to the mixed-species nature of the culture, of the exact total of reads belonging to each species.

The results in Table 2.6 indeed confirm that Sargasso performs an almost proportionally identical separation on the sub-sampled rat reads compared to the full-size sample. Once again, but now with experimental data, we can see that Sargasso mis-assigns very few reads (0.16%). Indeed its performance is worse than for the simulated data in 2.1, with a lower number of reads assigned to the species of origin and a marginally higher amount mis-assigned, however this is to be expected with real as opposed to simulated biological noise.

Figure 2.12 shows that the results for the experimental RNA-Seq closely mirror those of our 50bp simulated dataset for the rat. The variation of the key assignment variables for the experimental data cause changes to the F1 score of a similar magnitude and proportion to those observed for the simulated datasets. That the F1 scores are only marginally lower demonstrates that Sargasso proves almost equally effective on real RNA-Seq data as on simulated data.

The results from the mixed-species ADcc dataset (replicate 1 of the BiC condition), Table 2.6, show a small portion of reads were rejected due to species ambiguity: 1.9% for the rat and 2.6% for the mouse. Whilst the number of ambiguous reads is the same amount for both species, the number of total reads that map to the mouse genome is lower as a result of there being fewer astrocytes than neurons in the co-culture. That the number of ambiguous reads is this low however is reassuring for the application of Sargasso to other closely related species. That the majority of reads that mapped to rat were assigned to rat (75.3%), yet the majority of reads that mapped to mouse were rejected (52.7%) is a disparity arising from the differing number of each cell-type in the co-culture.

2.3.9.1 *Separation of the Full ADcc Dataset*

Having demonstrated the separation of two samples from the AD dataset, I then used Sargasso to disambiguate the data from all 12 mixed-species samples, mapping to the genomes of both mouse and rat.

For the AD dataset, I used Sargasso to disambiguate reads from mouse astrocytes and rat neurons. The dataset consisted of 3 conditions each with 4 replicates, with rat neurons and mouse astrocytes present in each condition and replicate. The assignment figures for the separation can be seen by replicate in Table 2.7, with figures for the separation of protein coding reads in Table 2.8.

We can see that there is a higher number of reads from the rat neurons in the dataset, across conditions and replicates, with roughly 2 to 5 times more reads being assigned to the rat than to the mouse equating 75.2% to 82.5% of mapped reads. Conversely we see between 27.5% to 44.8% of reads mapped to mouse assigned to mouse. This disparity of assignment is likely reflective of the cell-type populations in the ex-

perimental co-cultures, with the higher number of reads mapping to mouse a result of rat reads mapping to the mouse genome, which can be seen in the higher proportion of rejected reads for the mouse than the rat. Unfortunately only limited cell purity testing was carried out and so cannot provide additional validation for these findings. We see a consistent level of ambiguity across the conditions with a mean of 2.74% or reads mapped to mouse and 1.91% of reads mapped to rat unassigned as a result of species similarity.

The protein coding reads in Table 2.8 were generated as in Section 2.3.8. We can see that Sargasso's performance on only reads from protein coding genes is not hindered by a higher degree of conservation, with a percentage loss approximately equal to that for all reads. This is in agreement with the results for the protein coding read separation for the OSm-MA dataset in Table 2.5.

ADcc 'conservative'		Mouse				Rat			
Condition	Replicate	Mapped	Assigned	Rejected	Ambiguous	Mapped	Assigned	Rejected	Ambiguous
Control (w/ TTX)	1	107566855	37092624 **(34.5%)**	68094229 **(63.3%)**	2380002 **(2.2%)**	143570935	84756928 **(59.0%)**	56434005 **(39.3%)**	2380002 **(1.6%)**
Control (w/ TTX)	2	69523733	14781855 **(21.3%)**	53029618 **(76.3%)**	1712260 **(2.5%)**	114499510	73425086 **(64.1%)**	39362164 **(34.4%)**	1712260 **(1.5%)**
Control (w/ TTX)	3	111421040	34346441 **(30.8%)**	74465109 **(66.8%)**	2609490 **(2.3%)**	159347528	97939574 **(61.5%)**	58798464 **(36.9%)**	2609490 **(1.6%)**
Control (w/ TTX)	4	74915044	22668031 **(30.3%)**	50464013 **(67.4%)**	1783000 **(2.4%)**	108565183	67910106 **(62.6%)**	38872077 **(35.8%)**	1783000 **(1.6%)**
Bicuculine	1	105999393	37598790 **(35.5%)**	66042894 **(62.3%)**	2357709 **(2.2%)**	141633484	84118511 **(59.4%)**	55157264 **(38.9%)**	2357709 **(1.7%)**
Bicuculine	2	85170094	19231262 **(22.6%)**	63785221 **(74.9%)**	2153611 **(2.5%)**	140810255	91630443 **(65.1%)**	47026201 **(33.4%)**	2153611 **(1.5%)**
Bicuculine	3	92157237	29231585 **(31.7%)**	60812923 **(66.0%)**	2112729 **(2.3%)**	131516688	80245006 **(61.0%)**	49158953 **(37.4%)**	2112729 **(1.6%)**
Bicuculine	4	73376721	24594146 **(33.5%)**	47099911 **(64.2%)**	1682664 **(2.3%)**	102069648	62079961 **(60.8%)**	38307023 **(37.5%)**	1682664 **(1.7%)**
Bicuculine+TBOA	1	84228968	30137052 **(35.8%)**	52207822 **(62.0%)**	1884094 **(2.2%)**	111951585	66655607 **(59.5%)**	43411884 **(38.8%)**	1884094 **(1.7%)**
Bicuculine+TBOA	2	81256071	21362312 **(26.3%)**	57826128 **(71.2%)**	2067631 **(2.6%)**	125313125	80407441 **(64.2%)**	42838053 **(34.2%)**	2067631 **(1.7%)**
Bicuculine+TBOA	3	100130376	31793624 **(31.8%)**	66086064 **(66.0%)**	2250688 **(2.3%)**	142757122	87322344 **(61.2%)**	53184090 **(37.3%)**	2250688 **(1.6%)**
Bicuculine+TBOA	4	82202146	26842312 **(32.7%)**	53232012 **(64.8%)**	2127822 **(2.6%)**	115704331	72507330 **(62.7%)**	41069179 **(35.5%)**	2127822 **(1.8%)**

Table 2.7: **Sargasso Species Assignment Summary for the AD Dataset**.
Sargasso is capable of separating multi-species experimental datasets, demonstrated here with the ADcc dataset. The total number of reads mapped to the genome of each species, mouse and rat, is displayed and then broken down into the number assigned, rejected and unassigned due to ambiguity which are listed for each species by condition and replicate. Sargasso was run with the 'conservative' filtering strategy.

ADcc 'conservative'		Reads Mapped to Genome		Reads Assigned		% Lost	
Condition	Replicate	Mouse	Rat	Mouse	Rat	Mouse	Rat
Control (w/ TTX)	1	88491963	96359140	32714327	60972234	63.0	36.7
Control (w/ TTX)	2	57702998	77313875	13127544	53543751	77.3	30.8
Control (w/ TTX)	3	92173370	106730734	30260222	70291692	67.2	34.1
Control (w/ TTX)	4	62619129	72784041	19927685	48502927	68.2	33.4
Bicuculine	1	88228447	94405796	32696593	59119118	62.9	37.4
Bicuculine	2	71183555	93119358	16775314	64827702	76.4	30.4
Bicuculine	3	76836105	87717739	25698141	56713317	66.6	35.4
Bicuculine	4	61356985	68376186	21717647	44050284	64.6	35.6
Bicuculine+TBOA	1	70379569	75548422	26300009	47343564	62.6	37.3
Bicuculine+TBOA	2	68099296	85122902	18893400	58510118	72.3	31.3
Bicuculine+TBOA	3	83428538	96181744	27923468	62600060	66.5	34.9
Bicuculine+TBOA	4	68760481	78590612	23712747	52460911	65.5	33.3

Table 2.8: **Sargasso Assignment of Protein Coding Reads for the ADcc Dataset**. Sargasso's 'conservative' strategy assigns protein coding reads for the ADcc with slightly less proportionate loss that seen for all reads in Table 2.7. The figures here describe uniquely mapping reads only, with reads lost here including both rejected and ambiguous reads.

When we analyse the gene expression for the ADcc-MA dataset downstream however we can see, in Figure 2.13A and B, that the astrocytic expression profile is better than expected for an *in vitro* culture, where cells regularly do not achieve the extent of behaviour and function of the same cell-types *in vivo*, and actually approximates the profile we would expect from astrocytes *in vivo*. From this we can conclude that the presence of neurons in *in vitro* co-culture allows for greater development and maturation of astrocytes. The analysis of the astrocyte expression profile thus demonstrates the utility for Sargasso in the study of NCAE.

This downstream analysis of the Sargasso separation also enabled the discovery of a wide programme of neuron-induced gene expression in astrocytes, as can be seen in Figure 2.14 and in our paper (Hasel et al., 2017), with hundreds of genes shown to be regulated by synaptic activity. These include genes relating to glutamate metabolism, in keeping with the experimental stimulus of the AD dataset used. Known astrocytic specific marker genes S100b, Aldh1l1 and Gfap were not affected as one might expect, given their cellular specificity, however interestingly Aqp4 which also belonged to this group was. In addition to observing the magnitude of NCAE on astrocyte gene expression, as seen in Figure 2.14, we also confirmed Notch signaling to be a key mediator of this transcriptional change, where it facilitates glutamate uptake in astrocytes in addition to driving and maintaining maturity in these cells (Hasel et al., 2017). cAMP/PKA-dependent CREB activation was shown to be a mechanism by which astrocyte gene expression was regulated by synaptic activity, including components of the astrocyte-neuron lactate shuttle through which pyruvate is converted

to lactate in astrocytes and exported to neurons to be used as a substrate. Overall NCAE from neural activity seem to underpin metabolic cooperation between neurons and astrocytes (Hasel et al., 2017).

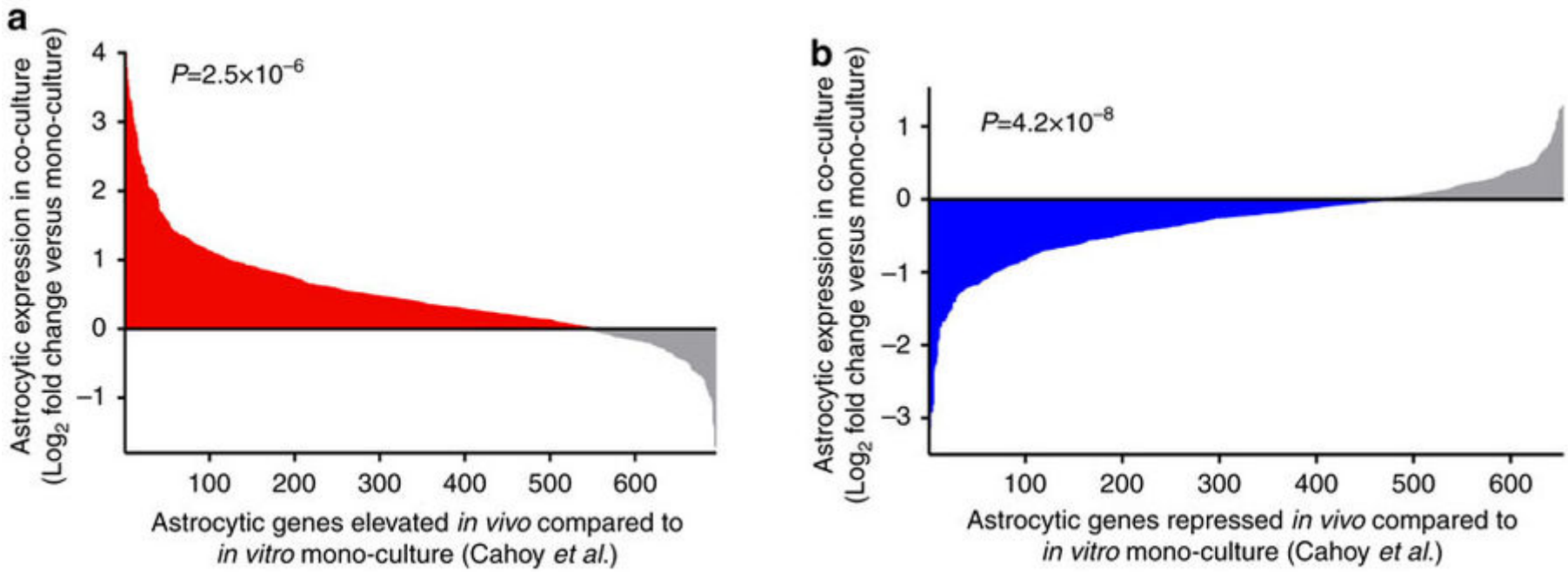
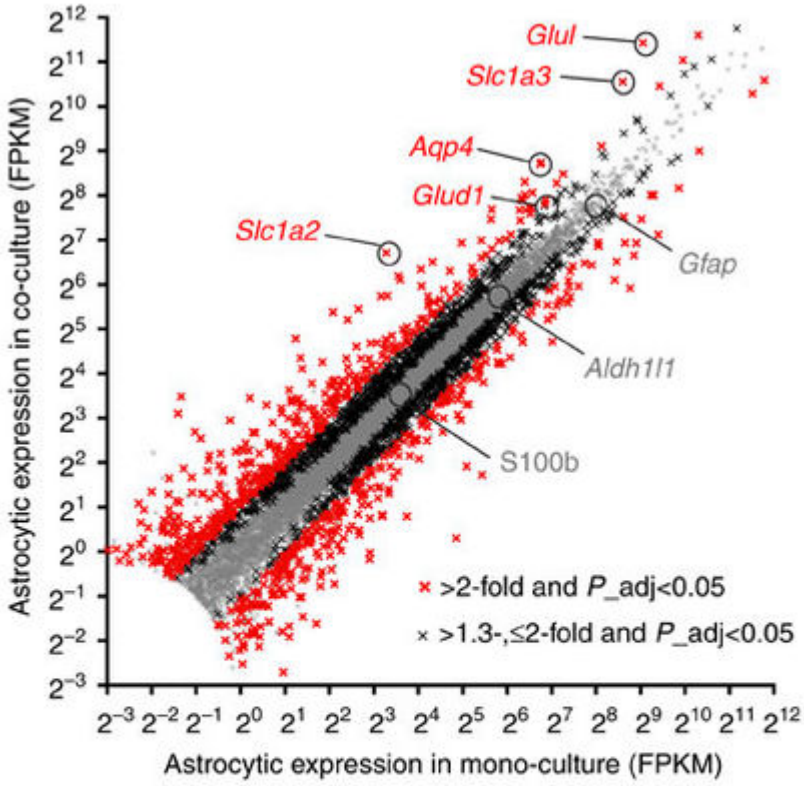


Figure 2.13: **Astrocytic Gene Expression Profile *in vitro* Versus *in vivo***. The *in vivo* astrocytic expression profile is here seen to be replicated in the majority of genes *in vitro,* shown here in the expression levels for the *in vivo* astrocytic gene expression profile identified in (Cahoy et al., 2008) as derived from the Sargasso separated mouse reads from the ADcc dataset. (A) Shows the gene expression for those genes found to be up-regulated by (Cahoy et al., 2008) *in vivo* with (B) showing the expression for those genes found to be down-regulated by (Cahoy et al., 2008). Figure reproduced from (Hasel et al., 2017).



	$Astrocyte_{Mus}$ mono-culture	$Astrocyte_{Mus}$ - $Neuron_{Rat}$ co-culture
Avg no. of reads mapped to mouse genome per sample (*n*=3)	38,056,262	13,380,921
% of mapped reads attributed to the mouse genome	99.9	12.4
Genes expressed >0.5 FPKM	13,122	13,163
Differentially expressed genes (≥2-fold difference)	763	
Differentially expressed genes (≥1.3-fold difference)	2,116	

Figure 2.14: **Change in FPKM of Astroycte Genes as a Result Co-culture with Neurons**. The abundance of astrocytic gene expression, in FPKM, can be seen to change greatly as a result of NCAE from co-culture with neurons. The FPKM from the co-cultured astrocytes can be seen here plotted against the FPKM of the astrocyte monoculture. Labelled genes are known astrocyte marker genes (S100b, Aldh1l1, Gfap, Aqp4), genes involved in glutamate metabolism (Slc1a2, Slc1a3, Glud1, Glul) or genes previously identified as regulated by synaptic activity (Slc1a2, Slc1a3). Summary statistics are listed with regard to differential expression. Figure reproduced from (Hasel et al., 2017).

2.3.10 *Comparison to Existing Methods*

Whilst Sargasso is the only *in silico* tool that has been designed for the general purpose separation of mixed-species RNA-Seq by species, there are two other methods developed for the separation of mixed-species RNA-Seq reads from xenograft samples. These tools, 'Xenome' and 'Disambiguate', are designed to separate 'graft' reads from those belonging to the 'host' species, usually human for the former and mouse or rat for the latter. In these instances the majority of the reads will usually belong to the host. As they are designed to separate graft from host, these tools can only separate data containing two species.

Disambiguate maps reads to both species genomes using the `TopHat` or `STAR` read aligners, for the latter it then assigns them by comparing their 'alignment score' (AS) and using their edit distance to break ties when AS is equal for both species (Ahdesmäki et al., 2016). We have used Disambiguate with `STAR` both as this is most comparable and `TopHat` is now deprecated.

Despite the specific use-case of these two tools, their methodology allows a user to apply them in a more generalisable manner if they wished. As a result, and considering them to be the closest tools for direct comparison, we applied both tools and Sargasso to the same single species RNA-Seq datasets in order to evaluate performance for general purpose disambiguation of mixed-species RNA-Seq. The datasets used were the 1N1 OSm-MN sample (mapped to mouse and rat), the RatOnly ADm-RN sample (mapped to mouse and rat) and the publicly available datasets SRR1930152 (only mouse reads; mapped to mouse and human) and SRR387400 (only human reads; mapped to human and mouse). The former two were chosen as they contain data from our own datasets of the two evolutionary close species Sargasso has been tested with thus far and the latter two as they were the data used in Disambiguate's publication (Ahdesmäki et al., 2016).

When attempting to apply Xenome to this data, we were unable to compile its k-mer reference for the rat genome. As such we presume that its authors did not intend for its use outside of human-mouse xenograft studies and as such we were not able to perform a comparison. Given this limitation we did not further apply the tool as this precludes Xenome's use as a general purpose method for the separation of mixed-species RNA-Seq.

The thresholds for the assignment variables used by Sargasso can be seen in Table 2.9. The results of Sargasso, for each of its filtering strategies, and Disambiguate can be seen in Table 2.10. As Disambiguate has a single fixed strategy for assignment, we could not perform repetitions to better compare with each of Sargasso's filtering strategies.

Tool	Disambiguate	Sargasso "conservative"	Sargasso "best"	Sargasso "recall"	Sargasso "permissive"
Mismatch threshold	N/A	0	1	2	25
Minmatch threshold	N/A	0	2	10	25
Multimap threshold	N/A	1	999999	999999	999999
Overhang threshold	N/A	5	5	5	0
Reject multimaps	N/A	TRUE	FALSE	FALSE	FALSE

Table 2.9: **Summary of Variable Thresholds for Each Assignment Strategy in the Existing Methods Comparison**. This table summarises the variable thresholds for the Sargasso filtering strategies used in the comparison to Disambiguate in this section. Disambiguate has no functionality for users to alter its assignment strategy, which does not use these variables, and so no value can be placed for them here.

From the assignment results we can see that the performance of the methods does not change substantially with the sample used. Disambiguate clearly implements a 'permissive' assignment strategy that prioritises obtaining the maximum number of reads assigned to the correct species over minimising mis-assignment. Disambiguate can be seen to clearly assign more reads to their species of origin than Sargasso for every filtering strategy except the 'permissive' strategy, however we can also see that Disambiguate also incorrectly assigns more reads than Sargasso for every filtering strategy except the 'permissive' strategy, indeed mis-assigning 4.75% of mapped reads for the ADm-RN sample. Overall we see a mean of 3.2% of mapped reads mis-assigned by Disambiguate compared to a mean of 0.11% for Sargasso's 'conservative' strategy.

The stringent thresholds of Sargasso's 'conservative' strategy can be seen to significantly impact the proportion of reads assigned to the species of origin compared to both Disambiguate and any of Sargasso's other strategies, with a mean of 51.2% of mapped reads assigned to the species of origin for Sargasso compared to a mean of 98.5% for disambiguate, as a result however we do see very few reads mis-assigned.

Sample: SRR1930152		**Origin Species: Mouse**			**Other Species: Human**		
Method	Total Reads	Reads Mapped %	Assigned % Total	Assigned % Mapped	Reads Mapped %	Assigned % Total	Assigned % Mapped
Disambiguate	24056144	96.47	96.139	99.653	12.494	0.330	2.644
Sargasso "conservative"	24056144	96.712	29.256	30.250	12.494	0.008	0.065
Sargasso "best"	24056144	96.712	80.540	83.279	12.498	0.209	1.669
Sargasso "recall"	24056144	96.712	88.727	91.743	12.498	0.240	1.921
Sargasso "permissive"	24056144	96.712	95.126	98.360	12.498	0.296	2.368
Sample: SRR387400		**Origin Species: Human**			**Other Species: Mouse**		
Disambiguate	59653070	83.395	83.290	99.873	10.068	0.157	1.555
Sargasso "conservative"	59653070	83.418	32.625	39.110	10.083	0.002	0.024
Sargasso "best"	59653070	83.418	57.576	69.021	10.083	0.021	0.205
Sargasso "recall"	59653070	83.418	66.148	79.297	10.083	0.034	0.332
Sargasso "permissive"	59653070	83.418	79.247	95.000	10.083	0.150	1.486
Sample: 1N1 (OSm-MN)		**Origin Species: Mouse**			**Other Species: Rat**		
Method	Total Reads	Reads Mapped %	Assigned % Total	Assigned % Mapped	Reads Mapped %	Assigned % Total	Assigned % Mapped
Disambiguate	51996498	94.835	91.905	96.911	44.292	1.651	3.728
Sargasso "conservative"	51996498	95.115	61.845	65.021	44.615	0.086	0.192
Sargasso "best"	51996498	95.115	83.790	88.093	44.615	1.388	3.111
Sargasso "recall"	51996498	95.115	88.734	93.291	44.615	1.715	3.843
Sargasso "permissive"	51996498	95.115	92.263	97.001	44.615	1.844	4.133
Sample: ADm-Rn		**Origin Species: Rat**			**Other Species: Mouse**		
Disambiguate	48108495	95.416	92.936	97.401	44.060	2.092	4.749
Sargasso "conservative"	48108495	95.590	67.271	70.374	44.135	0.061	0.137
Sargasso "best"	48108495	95.590	85.886	89.848	44.135	0.541	1.226
Sargasso "recall"	48108495	95.590	89.656	93.792	44.135	0.782	1.772
Sargasso "permissive"	48108495	95.590	92.918	97.205	44.135	1.011	2.291

Table 2.10: **Read Assignment for Sargasso and Disambiguate Over 4 Samples.** Disambiguate can be seen to assign reads with comparable accuracy to Sargasso's permissive strategy. That these approaches both result in a high level of false positives will have a negative impact on downstream analysis, something that can be seen to be minimised through Sargasso's more stringent filtering strategies. This table details the quantity of reads assigned to the species of origin and to the other species mapped for Disambiguate and Sargasso for each of its four filtering strategies. The number of reads mapped to each species is reported in addition to the number assigned to each species both as a percentage of reads mapped and of total reads. Two samples from our experimental data were used in conjunction with the two used in (Ahdesmäki et al., 2016).

2.4 DISCUSSION

Accuracy of Separability & Data Loss

With this project, we set out to investigate whether it was possible to separate mixed-species RNA-Seq reads *in silico*. In Section 2.3.1 we demonstrated, from a simulated dataset of all theoretically possible reads for the rat, that, by using several key variables taken from read alignment to a genome, we could correctly attribute reads to their species of origin in a single species scenario. This was shown to be possible with a high level of accuracy, with only a minority of genes losing more than a few percent of their reads, even when using the most conservative thresholds for the key variables. Whilst assignment precision was high, the recall achieved was lower, with many reads being rejected either due to species similarity between rat and mouse or due to the incomplete nature of the rat genome. A surprisingly small percentage of reads, only 0.88%, were declared ambiguous due to equal mapping to both genomes. This accuracy of assignment was seen for both short (50bp) and long (150bp) reads with and without simulated sequencing error. That we could achieve such high precision and F1 scores for simulated data was both promising and unexpected; given the small genetic distance between rat and mouse we expected a higher degree of uncertainty.

When we look at the data that was rejected, only a small portion of it is due to sequence ambiguity between species. As such the rest must be lost either as a result of sequencing error, or simply due to poor mapping, either as a result of an imperfect portion of the reference genome, or due to exceeding the thresholds for the key assignment variables; in Section 2.3.7.1 we demonstrated that this was indeed the case. When conservative variable thresholds were relaxed we saw a decrease in unmapped reads from 8.4% to 3.7% for mouse and 13.5% to 5% for rat. The number of genes losing a substantial ($>$ 20%) portion of their reads due to Sargasso also fell by 13.2% to 1.1% for mouse and by 19% to 1.6% for the rat. Thus we can see that stringent thresholds, whilst ensuring greater certainty in practice, do result in a significant loss of data. We can also indeed see the impact of genome quality, with the mouse results being stronger than the rat. When we take into account approximately 1% of reads lost to sequence ambiguity, we can thus see that our false negative rate, the amount of reads incorrectly rejected, falls to less than 5% for both species when using the relaxed thresholds in our 'best' filtering strategy. We can see these reads lost are also dispersed amongst genes, with only 1.1% of mouse and 1.6% of rat genes losing $>$ 20% of reads. These percentages are reduced further by assigning these reads to features, as would be done in a standard downstream analysis, giving us 0.4%

and 1.4% for mouse and rat respectively for our 'best' strategy, with our stringent 'conservative' strategy achieving 5.2% and 5.4% respectively.

Through further investigation we revealed that, for our 'conservative' strategy, the genes whose reads have only a single mapping to their species of origin but who have multiple mappings to the other genome are those most likely to lose a large proportion of their reads. We also showed that the genes that experience this read loss are those that are more likely to be conserved or similar between the species we are considering, something which Sargasso cannot mitigate.

Whilst it is a much smaller occurrence than incorrect rejection, Sargasso does incorrectly assign a small minority of reads. In Section 2.3.7.2 we demonstrated that this proportion of reads is negligible for conservative thresholds, at 0.002% and 0.001% genes with $> 1\%$ incorrect reads for mouse and rat, however when relaxed under the 'best' strategy we do see this rise to 0.9% for the mouse and more concerningly 2.6% for the rat, though only 0.2% of genes observe $> 20\%$ incorrect assignment. Again this disparity between species may be a result of genome quality.

These results therefore demonstrate that whilst Sargasso may lose a small amount of data when using stringent thresholds, we should not expect the use of Sargasso to introduce significant bias in downstream analysis. Indeed much of the data that is lost is due to species similarity. That we do see a rise in incorrect assignment when relaxing these thresholds is an issue that could have a small impact on downstream analysis, however these are minor in comparison to the bias and noise introduced by physical separation techniques. This is well illustrated in Section 2.3.6, and in our paper (Hasel et al., 2017), where we demonstrated the significant impact of an imperfect physical separation on gene expression where there is 5% contamination with incorrect cells.

When the data used was restricted to reads from protein coding gene only, in Section 2.3.8 and for the AD co-culture in Section 2.3.9.1, we can see that Sargasso's performance is no worse than when applied to all reads. Indeed the proportion of protein codes reads lost as a result of rejection and ambiguity can be observed to be marginally less for many samples. As a result we can see that, whilst genes of very high cross-species conservation present a difficulty for Sargasso's methodology, we do not see a higher proportion of rejection when only reads from protein coding genes are considered. Thus it is likely that only a smaller portion of protein coding genes are of sufficiently high sequence conservation to present an obstacle for Sargasso, which is supported by our findings on theoretical reads in Section 2.3.7.1.

Separability Across Greater Genetic Distance and of Multi-species RNA-Seq Data

When we expanded our investigation to trial separations using different species, in Section 2.3.3, the results for the simulated mouse reads were even marginally better. Given that the mouse genome is more complete and better annotated than the rat, this is the likely cause of the minor disparity. Read assignment for more genetically distant species, such as human and zebrafish, also achieved a similar level of assignment accuracy with human being the most separable, when separated against the rat. That zebrafish was less accurately separable that human was a slight surprise given the greater genetic distance and thus greater genomic dissimilarity, however, again, its genome is less complete and well annotated than certainly that of the mouse and human.

Having seen such promising separability for the simulated data and across genetic distance, the next step was to apply Sargasso to real experimental data, as described in Section 2.3.9. The similarity in Sargasso's performance for the experimental rat versus the simulated rat dataset when both were separated against the mouse, in Figure 2.12, in terms of F1 score was very close with only marginally lower (between roughly 0.01 and 0.02) F1 for the experimental data. This is encouraging for the interpretation of the other simulated results. For the mixed-species mouse and rat experimental dataset we saw a different level of mapping that is reflective of the different cell populations, with a heightened level of ambiguity compared to the single species data. Whilst the explicit accuracy cannot be confirmed in a mixed-species case, that the expression profile for the mouse astrocyte reinforced the findings of previous studies (Hasel et al., 2017) indicates that the separability is indeed of high quality.

We have thus demonstrated that it is possible to accurately separate mixed-species RNA-Seq reads, simulated or experimental, *in silico* with only a minor sacrifice in terms of recall. With this separability, our method presents a novel step forward in enabling the study of non-cell-autonomous effects in *in vitro* cell cultures.

Effect on Downstream Analysis

Given that the use of Sargasso does impact the source data both through read loss, and with the mis-assignment of a minority of reads, the final area of performance evaluation for us to investigate was how our tool affected the downstream analysis of the data it was separating. As such we used a standard differential expression analysis, using DESeq2, as described in Section 2.3.7.3.

Having investigated the effect of Sargasso, using both 'conservative' and 'best' filtering strategies, against a standard genome mapping with `STAR` for single species data we could see that the impact on both up-regulated and down-regulated dif-

ferentially expressed genes was small. For the up-regulated genes there was only 104/3714 genes (2.8%) that were not discovered by either Sargasso strategy, for the down regulated there was 102/3376 genes (3%). When we look at Sargasso's performance by filtering strategy we can see that these figures rise marginally for 'best' to 114/3714 genes (3.1%) and 110/3376 genes (3.3%) and significantly for 'conservative' to 241/3714 genes (6.5%) and 306/3376 genes (9.1%). As a result, the choice of thresholds has a notable impact on the downstream analysis, with the 'conservative' filtering strategy resulting in a substantial reduction in the detection of differentially expressed genes.

Whilst the main effect of Sargasso application on the data is this loss of previously differentially expressed genes, we do see a small number of newly differentially expressed genes. For up-regulated genes for both Sargasso filtering strategies we see 50/3714 genes (1.3%) newly differentially expressed and for down regulated we see 27/3376 genes (0.8%) for 'conservative' and 34/3376 genes (1%) for 'best'. While these totals are much lower than the number of previously differentially expressed genes lost by Sargasso, it is necessarily something that must be borne in mind when assessing results downstream.

With the application of Sargasso to experimental data, in Section 2.3.9, we analysed the astrocyte expression profile in order to investigate neuronally stimulated transcriptomic change. In doing so we were able to confirm the result of previous findings regarding the astrocytic expression profile and therefore demonstrate that Sargasso's separation enabled the investigation and confirmation of a NCAE. Indeed that we could confirm similarity between *in vitro* and *in vivo* astrocyte expression at all was a key finding of our paper: (Hasel et al., 2017).

So whilst Sargasso does have an impact on the data it separates, we have shown in Section 2.3.7.1 that this is limited to a minority of genes with high level of sequence conservation between the genomes of the two species being separated, indeed our results showed this affect could also be mitigated through the use of s less stringent filtering strategy. Again, the biggest issue appears to be more about mitigating data loss than preventing bias. Whilst we have again compared our performance against a standard genome mapping and differential expression, it is important to emphasise that this methodology is not possible for multi-species datasets or the study of non-cell autonomous effects. Indeed when we simulated the effects of contamination from a physical separation, as we did in Section 2.3.6 and in our paper (Hasel et al., 2017), we see a much worse performance with the FPKM of 863 genes changed greater than two-fold, for 216 of which we see a change greater than ten-fold. Thus whilst our method does affect the data, the magnitude of this effect is thus smaller than those risked by physical separation.

Applicability

Having demonstrated that Sargasso performs well for both genetically close and distant species and that this performance is transferable from simulated to real data it is important to discuss the applicability of the tool.

When considering species for Sargasso's use there are several key considerations to be made. Firstly, and perhaps most importantly, is the quality of a species' reference genome and annotation as any gaps or poorly annotated regions may lead to mis-assignments or rejections and thus impact downstream analysis. Secondly is whether the species selected genetically close enough for the derived cell-types to approximate the behaviour of cells from the same species when co-cultured *in vitro*, at the very least with regard to the functions being investigated. With improving genomes for laboratory mouse strains, if Sargasso proves these separable, this may be the best avenue in future for assuring cellular compatibility.

Sargasso presents a novel way forward for the study of non-cell-autonomous phenomena which we have demonstrated through our group's own research. We used Sargasso to identifying neuron dependent gene-expression in astrocytes (Hasel et al., 2017), where we discovered the breadth of the neuron-regulated genes in the astrocytic transcriptome, in particular those regulated by synaptic activity via mechanisms involving cAMP/PKA-dependent CREB activation, and the importance of Notch signalling in the neuron mediated maturation of astrocytes, in addition to providing further evidence of astrocyte-neuron metabolic cooperation via the astrocyte-neuron lactate shuttle. We have also have a protocols paper in press demonstrating how to co-culture mixed-species cells and apply the Sargasso method for the purpose of studying non-cell-autonomous phenomena (Qiu et al., 2018).

We have also used Sargasso to confirm species specific functional differences. Sargasso was applied to study evolutionary divergence in the activity dependent gene expression of developing neurons (Qiu et al., 2016). Our paper compared the transcriptional response in mouse and human neurons before using Sargasso to confirm species specific gene responses from the Tc1 transchromosomic mouse strain containing human chromosome 21. This allowed us to study the response of orthologous genes in the same cellular environment and thus confirm that observed behavioural difference must result from difference in genetic sequence.

Whilst the latter example has somewhat limited scope, requiring the use of very specific mouse strains, the former methodology can be extended to include additional species specific cell-types or easily generalised to investigate the interactions of different cell types. Indeed, it is possible that even two different types of tissue could be studied in this manner. As such there is an incredibly wide area for potential ap-

plication of our tool for studies of this manner. It is therefore our hope that other laboratories will build upon our work and use our methodology and Sargasso to further characterise cell behaviour through the study of non-cell-autonomous effects.

Comparison to Existing Methods

In Section 2.3.10 we compared the performance of Sargasso to the closest applicable tool, Disambiguate. Over the four samples to which they were applied, two from our datasets and two from the Disambiguate paper (Ahdesmäki et al., 2016), we saw each method perform with reasonable consistency. Disambiguate, as a result of being developed for xenograft studies that primarily use the more distant species human and mouse, implements a fixed 'permissive' strategy that prioritises maximal read assignment to the species of origin over minimising mis-assignments. As a consequence whilst we see almost all reads mapped to the species of origin assigned to it, we see a high level of mis-assignment in the Disambiguate results, with a mean of 3.2% mis-assigned over the four samples tested. As we have shown in Section 2.3.6, this will have a significant negative impact on downstream analysis.

Whilst Disambiguate clearly outperforms Sargasso's 'conservative' filtering strategy, with regard to proportion of reads assigned to species of origin, this difference in performance is smaller to the 'best' and 'recall' strategies and only marginally different from the 'permissive strategy'. As a result the flexibility of Sargasso can be user adjusted to be both replicate the performance of Disambiguate and also to be best appropriate for the experimental data being separated, in terms of species distance and data quality. It is also of worthy note that Disambiguate's assignment strategy makes use of the 'alignment score' (AS) criteria produced by the `STAR` read aligner, whose inappropriateness we have previously highlighted in Section 2.2.2.1.

If we are to consider the general purpose case for mixed-species *in silico* read assignment, for which Sargasso was designed, the arguable priority is the minimising of false positive assignments. As if sufficiently deep sequencing is used to produce the data then we do not need to map all of the reads to get a full picture of gene expression as there will be sufficient redundancy, so long as there is no bias in assignment. As we have shown in this chapter, the effect of Sargasso downstream is minimal. Disambiguate, due to its permissive strategy, has a very high number of false positive assignments. For its designed use-case in separating reads in graft studies, the priority is to map as many reads from the graft to the species of origin as there will necessarily be fewer reads than for the host species and these are the reads of primary interest; thus a higher false positive rate is tolerable. For a mixed species dataset with a less extreme distribution of reads for each species, we do not have

to sacrifice precision for recall in such a manner as there will be enough reads from each species, if sequenced to sufficient depth, to achieve a full picture of gene expression. We are also likely to be equally interested in reads belonging to both species. Thus as Sargasso's performance can be tuned to reflect the experimental conditions, with the ability to prioritise minimising false positives or to maximise read assignment or indeed any degree in between, it is more applicable to the general case than Disambiguate. Whilst Disambiguate performs marginally better for its intended use-case than the permissive Sargasso, the fact that it uses AS to assign reads reduces confidence in its assignment as this metric is not comparable across species, due to difference in intron lengths between species which forms part of the score, and thus inappropriate for the purpose in which it is employed.

As a result we can conclude that Sargasso's performance can be comparable if required, due its flexibility, however is more widely applicable, more user friendly and can be used to disambiguate reads from mixed-species datasets containing more than two species.

2.5 FUTURE WORK

Whilst Sargasso is a published and publicly available tool, there is still much work that can be done to improve its function. In this section I will describe several such improvements that the tool would potentially benefit from.

2.5.1 *Pre-processing to Identify and Remove Conserved Regions Between Species*

Sargasso takes as input the reference genome and annotation for all species in a multi-species dataset. Depending on the genetic distance between these species and the lengths of the reads, there will be highly conserved regions whose sequence is identical between the genomes and thus whose reads cannot be correctly attributed by Sargasso. Whilst this is a situation that cannot be avoided, though whose effect may be mitigated to some extent by using long reads, these regions could be excluded from, or simply flagged within, the analysed genome in order to reduce processing time.

The identification of this 'core transcriptome' conserved between species would only have to be updated with each genome release, rather than through each application of Sargasso, thus potentially saving processing time in subsequent executions through reduction of comparable genome length. This would best be calculated in a pair-wise manner between species. The size of the 'core transcriptome' will vary necessarily based both on genetic distance of the species used and the length of reads used, and as a result it will be most beneficial for shorter read lengths and closely related species as there will be more inseparable locations and more so for shorter stretches.

A study on the exact size of the inseparable transcriptome between species for different read lengths will thus be the first step in determining whether the time saved by precluding its use is worthwhile, as only a minority of two species' trancriptome will be conserved without variation. For example a previous study found only 481 regions of perfect sequence shared between human mouse and rat that were 200bp or longer (Bejerano et al., 2004).

2.5.2 *Additional Genome Library Selection*

Whilst Sargasso takes as input the reference genome and annotation for each species, this genome is not a complete set of sequence information for any species. There are several separate small libraries of specific sequence information that can be added in,

such as ribosomal rRNA for example, that can improve the accuracy of Sargasso's assignment. Whilst these sequences could presently be manually combined by a knowledgeable user, the inclusion of execution flags to enable the specific and intentional incorporation of these additional sequence libraries would improve Sargasso's usability.

2.5.3 *Separation Trial for Laboratory Strains of a Single Species*

As we have demonstrated, the mixed-species RNA-Seq of closely related species, such as mouse and rat, can be separated by Sargasso. However even with such closely related species, we cannot always expect generalised conservation of cell behaviour between all cell-types and for all experimental conditions, therefore ideally we would like to be able to use different laboratory strains of the same species in order to maximise certainty with regard to cell behaviour.

A separation conducted on Human and Chimp in (Qiu et al., 2018) added additional evidence that, whilst there was reasonable and expected data loss resulting from genome similarity, separation of species at a very close genetic distance and with sufficiently complete genomes is possible.

Whilst a trial by Owen Dando demonstrated that there may be enough difference between the genomes of different mouse strains for Sargasso to correctly assign reads, the reference genomes for these laboratory strains are not sufficiently complete or well annotated at present to enable an adequate trial of Sargasso at this time. The Mouse Genomes Project (Adams et al., 2015) at the Sanger Institute is presently working on and improving these genomes, so it is hoped that a Sargasso trial for strain-wise separation could take place in the near future.

If separation of mouse strains proves possible, subsequent analysis would need to be carried out to determine whether there would be sufficient genomic difference for experimental phenomena to be distinctly visible in downstream analysis of the transcriptome.

2.5.4 *Trial of Deep Learning Methodology for Species Assignment*

As detailed in Section 2.2.2.2, Sargasso uses an intuitive combination of variables derived from genome alignment in order to assign reads to a species of origin. This mechanism is reasonably simple and whilst its precision is very high its recall could likely be improved upon through the application of machine learning techniques.

Deep learning techniques are presently being applied with success to various problems in computational biology, from variant calling (Poplin et al., 2018) to classifi-

cation of metagenomic data (Fiannaca et al., 2018; Fioravanti et al., 2018). The latter is of particular interest as the problem is similar to our own: the assignment of sequence reads to species of origin. For our purposes classifiers could be trained on the genomes of the species in our data, as in this study (Fiannaca et al., 2018) classifiers are trained on taxon specific sequence data. Once trained these could be used either to provide supplementary assistance to or indeed a replacement for, depending on performance, our current assignment algorithm. Whilst ribosomal rRNA is likely not the best feature for our classification problem, different features could be investigated for use. A trial of deep learning methodology could thus prove a useful avenue for future investigation.

2.6 AVAILABILITY

Sargasso is publicly available on GitHub, at `https://github.com/statbio/Sargasso`, under open source license as a command line executable Python package. Instructions for installation and execution in addition to information on dependencies can be found in our online documentation at `https://statbio.github.io/Sargasso/`.

3

PATHWAY ENTROPY

3.1 MOTIVATION

The association of the information within biological knowledge repositories with experimental data is an ongoing challenge within computational biology. Our knowledge is incomplete and ever expanding and the data itself is inherently noisy. A common procedure for the linking of knowledge and data has long been to apply tools using frequentist statistics to determine whether a biological feature of interest is significantly over-represented in a set of co-expressed genes compared to what we would expect by chance, a process frequently termed 'enrichment' that has been discussed in more detail in Section 1.3.2.

Network modelling is an increasingly popular methodology in computational biology for the analysis of NGS data. For the analysis of gene expression data for example, use of network methods has shown a significant advantage over non-graph based approaches in uncovering features of interest (Langfelder and Horvath, 2008). Importantly the structure and topology of networks constructed from biological data have been used to uncover underlying functional properties (Pržulj et al., 2004). For further discussion of network biology and approaches see Section 1.3.1.

Biological pathways are collections of interacting genes, proteins and other biological molecules whose functional relationships have been experimentally verified or inferred from experimental data. They are manually curated repositories of biological knowledge relating to specific phenomena ranging from metabolic processes to disease. If the topology of networks constructed from gene expression data represent functional associations then it can be theorised that biological pathways, whose gene members functionally interact with one another, should represent strong sub-units within these networks. Tools such as "Weighted Gene Co-expression Network Analysis" (WGCNA) (Langfelder and Horvath, 2008) cluster gene co-expression networks on the basis that genes which are similarly co-expressed are likely to be functionally related. As such we might expect gene members from a biological pathway to be clustered together; however, given the size and range of gene function within these networks it is unlikely that this clustering would neatly partition pathway members from the wider noisy dataset. Therefore we need a means to confirm the association of pathways with clusters.

Despite the increased uptake of network methods for computational biology in the last decade, the methods for associating biological knowledge with these networks has neither significantly advanced nor adapted to the new sources of information present in these networks. Only one method that I have found attempts to use the structure of the network to inform pathway association: "Differential Network Analysis" (DiNA) (Gambardella et al., 2013). This method applies Shannon Entropy (Shannon, 1948) to a series of networks or sub-networks to derive a metric of pathway representation throughout the data. The fewer sub-networks that contain pathway members, the less disordered that pathway's representation is within the data and thus the smaller the entropy metric. Representation of a pathway's members over multiple sub-networks within a gene co-expression network demonstrates that pathway members are not co-expressed together and are thus in a state of disorder, resulting in a higher metric. Whilst this method proposes an interesting formula for analysing pathway involvement in gene co-expression networks, it does not make full use of the structural information available; increasing information content in this manner would seem an easy avenue for improving overall accuracy and performance. This method, though not widely used, therefore provides a good starting point and in this project I will be improving its methodology to better utilise structural information from the data and thus allowing for more informed enrichment of biological pathways for network based gene expression data.

In addition to the improvement of network based enrichment methodology, the resulting tool from this project, if sensitive enough, can also be applied in an evaluative capacity. Construction of gene co-expression networks requires making choices of several key network construction variables: pre-processing method, correlation method, edge weight thresholding method and clustering method as well as the parameters of that cluster method. Whilst there are many options for these choices, finding the most suitable for a given task is a non-trivial operation. Critical assessment of network construction methodology is a difficult task and there is not a purpose-built tool that exists for doing so, that I am aware of. I propose that the use of a network based pathway entropy tool, such as that I have outlined here, if suitably sensitive, could serve as means for measuring and evaluating change over a range of network construction variables. This can be done either by assessing the parameter choices that give the most biologically meaningful pathway enrichment or by most faithfully reproducing a strong artificial signal in the data. Such an application would thus allow for an informed variable choice when constructing gene co-expression networks.

3.2 IMPLEMENTATION & EXPERIMENTAL DESIGN

In this section I will describe the entropy methodology I have developed and how it has been applied to experimental data. I will justify the approaches taken, detail the tools and component methods I have used as well as any tested alternatives.

3.2.1 *Approach*

3.2.1.1 *Language*

I chose to develop this method in the R statistical programming language because of the libraries and tools the platform contains for computational biological analysis, both in its core repositories and in the Bioconductor (Gentleman et al., 2004) dedicated repository. Python was considered as it too has many tools for computational biology; however the availability of the WGCNA network construction tool in R was a key consideration.

That the method has been built entirely in R and does not require external tools or packages, beyond the R packages it makes use of, enables it to be easily packaged in the future for public distribution.

3.2.1.2 *Intended Use Cases*

This method is intended to be used for the analysis of NGS gene expression data. Whilst it has been developed for use with RNA-Seq data, data produced by older technologies, such as microarrays, should also be compatible. However as the network construction takes as input summarised read data, produced by a tool such as `featureCounts`, any biological data that can be processed into this form can be used. There is no specific use case with regard to the biological system under investigation.

3.2.1.3 *Choice of Biological Pathway Database*

There are several databases of biological pathways that are publicly or semi-publicly available. The two that were considered for this project were the KEGG Pathway Database and Reactome (Croft et al., 2010).

The KEGG Pathway Database is a semi-publicly available, manually curated resource that has a long history of use for pathway enrichment. Pathway data is extractable through the KEGGREST R package in the databases' 'KGML' format from which the gene interaction data, that constitutes a pathway's functional topology, can be derived.

Reactome is a newer pathway database and has a very different structure. Whereas KEGG contains a series of individual pathways, Reactome pathways are described in a hierarchical and modular fashion allowing for differentiation of more general and more specific components. It has an API through which data can be extracted and interaction data is easily obtainable given the database's modular structure.

Whilst many of KEGG's pathways also contain nested representations of other pathways, the structure of KEGG pathways is not implicitly modular or hierarchical as in Reactome. An example of this nesting is as follows; if a disease pathway impacts a certain metabolic pathway then the impacted portion of the latter pathway will be included (or 'nested') in the disease pathway.

Whilst Reactome has a higher level of detail available, I chose to use KEGG as its simpler structure makes it a more practical choice. Whilst Reactome may be desirable for the extra structural information it provides, its hierarchical structure would make it more difficult to work with in addition to the fact that as a more recent database than KEGG, its contents will have been subject to less evaluation.

3.2.2 *Description of Entropy Methodologies*

Entropy, as originally defined by Shannon (Shannon, 1948) in the field of information theory, has previously been applied to problems in the field of computational biology as discussed in Section 1.3.3. However its application to gene co-expression networks has been limited. In this section I will describe a key previous implementation before advancing to the alterations I have made to its methodology and describing the function of the resultant method through an example.

3.2.2.1 *A Definition of Entropy as Previously Applied in Network Biology*

The DiNA (Gambardella et al., 2013) implementation of entropy for the attribution of pathway involvement in a biological network is an application of information theoretic entropy to a problem of computational biology: determining the association, or enrichment, of biological terms to experimental data through examination of representation in a given set of genes. For a set of independent gene co-expression networks or sub-networks from a single network, this method calculates a single entropy metric for each pathway in a biological pathway database - KEGG for example. As my focus of application is on single clustered gene co-expression networks, I will henceforth describe the function of this tool within the context of evaluating the sub-network clusters from a single network. This entropy metric describes the level of disorder with regard to a pathway's representation in the data; a pathway whose members are represented in a single sub-network has an ordered representation, and

thus a low entropy, whereas a pathway whose members are dispersed across sub-networks has a disordered representation resulting in a high entropy.

The entropy, $H(V)$, for a pathway, V, is defined as:

$$H(V) = \sum_{i=1}^{N} P(V = i) \log\left(\frac{1}{P(V = i)}\right) \tag{3.1}$$

where $P(V = i)$ is defined as follows:

$$P(V = i) = \frac{n_i}{\sum_{j=1}^{N} n_j} \tag{3.2}$$

$H(V)$: The entropy value for pathway V. The larger the entropy value the smaller the likelihood that the pathway is involved with 0 indicating maximal involvement of the pathway in the data.

$P(V = i)$: The probability that the genes in pathway V are co-regulated only in the i^{th} sub-network.

N : The number of sub-networks we are investigating pathway co-regulation for.

n_i : The number of edges connecting nodes within the i^{th} sub-network whose genes are represented in pathway V.

3.2.2.2 *Modified Pathway Entropy Definition*

The following equations define our modified Pathway Entropy function. The modifications normalise for both the representation of the biological pathway, specifically the proportion of a pathway's gene members present in a sub-network, and the co-expression sub-network's connectivity, with regard to the proportion of sub-network connectivity belonging to genes that are co-expressed members of a pathway.

We normalise for pathway representation to control for situations in which strongly co-expressed partial representations of large pathways can lead to false positive enrichments. We thus place emphasis on whole pathway representation in the data.

The proportion of connectivity in the sub-network between pathway members is normalised for as, due to the noisy nature of biological data, pathway members will rarely be clustered in isolation from other genes and as a result it is important to take into account how well they are clustered. A sub-network where a majority of connectivity represents co-expression between pathway members demonstrates a stronger representation of a pathway than a large sub-network containing a majority of non-pathway connectivity where a pathway is incidentally well represented.

This variation utilises edge weights in place of a simple edge count to better preserve topologically relevant information and strength of gene associations.

The entropy for a pathway, V, is now defined as:

$$H(V) = \sum_{i=1}^{N} P(V = i) \log\left(\frac{1}{P(V = i)}\right)\left(\frac{(q - g_i)(f_i - n_i)}{q f_i}\right) \tag{3.3}$$

where $P(V = i)$ is now defined as follows:

$$P(V = i) = \frac{w_i}{\sum_{j=1}^{N} w_j} \tag{3.4}$$

$H(V)$: The entropy value for pathway V. The larger the entropy value the smaller the likelihood that the pathway is involved with 0 indicating clear involvement.

$P(V = i)$: The probability that the genes in pathway V are co-regulated only in the i^{th} sub-network.

N : The number of sub-networks we are investigating pathway co-regulation for.

n_i : The number of edges connecting nodes within the i^{th} sub-network whose genes are represented in pathway V.

g_i : The number of nodes within the i^{th} sub-network whose genes are represented in pathway V.

w_i : The sum of the edge weights for edges that connect nodes whose genes are represented in pathway V, within the i^{th} sub-network.

f_i : The number of edges connecting all nodes within the i^{th} sub-network.

q : The number of genes in pathway V.

I have implemented two different variations of the above equation: 'topology' (PE-T) which places emphasis on the representation of genes with functional annotation within a KEGG pathway and 'all' (PE-A) which emphasises all co-expression between all pathway gene members, including those whose role in pathway interactions is not explicitly annotated. Both of these methods are implemented as 'weighted' (PE-Tw/PE-Aw) and 'unweighted' (PE-Tu/PE-Au) in order to demonstrate the effect of using the edge weights as additional information; the unweighted implementation uses a simple edge count in place of the correlation values. Both methods are described in further detail below.

The Pathway Entropy 'topology' Implementation

In calculating the entropy, the Pathway Entropy 'topology' (PE-T) implementation takes only edges for w_i that exist between pathway members represented explicitly in a pathway's interaction schema, or topology; where functional interaction has been confirmed. This sub-group of genes is defined as that whose pairwise interactions can be extracted from the KEGG 'KGML' for each pathway by directly querying the KEGG database. These members represent the subset of the total members of the pathway for which interaction information is known and thus those annotated onto the KEGG pathway visualisations, an example of which can be seen in Figure 3.1. In order to best capture patterns of co-expressed activity, and likely pathway involvement, the PE-T approach considers all edges between genes within this subset. In doing so, all entropy thus calculated is explicitly based on the known subset of pathway genes for whom functional interaction has been confirmed and is explicitly defined in a given pathway. This adds an additional layer of information to the calculation which should result in a more robust entropy and an intrinsically representative enrichment.

As a result of limiting edges to this subset, we are taking a conservative approach to pathway enrichment that necessarily restricts the pathway information used to only that that is best characterised in the pathway database. Therefore for users who wish to identify new genes that may be involved in a pathway, this approach would be poor.

For the purposes of comparison, considering this is some departure from the original DiNA methodology, I have implemented a DiNA 'topology' adaptation which limits DiNA to only using pathway gene members within the same subset used by PE-T.

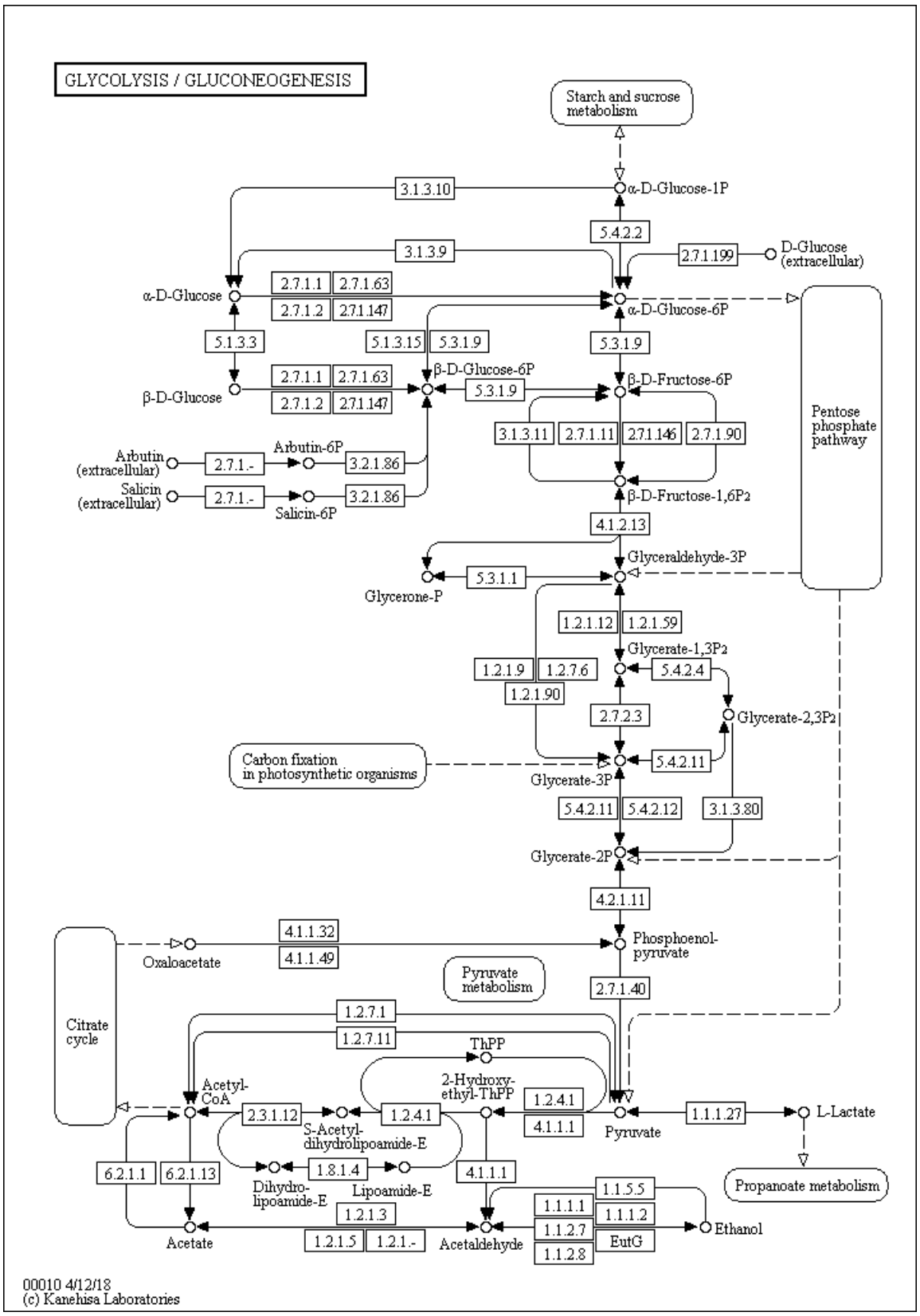


Figure 3.1: **KEGG Pathway Diagram 00010 - 'Glycolysis / Gluconeogenesis'**. This figure illustrates the structure of a KEGG pathway. The small circles represent biological molecules or compounds, with arrows representing direction of interaction, square numbered boxes represent pathway stages to which genes are annotated; it is these genes that form the subset used by PE-T. Rounded boxes represent where a pathway joins or interacts with a separate pathways in the database.

The Pathway Entropy 'all' Implementation

As not all interactions between genes in a given pathway are known or functional roles well characterised for the majority of biological pathways, the PE-T method is at the mercy of both established biological knowledge and pathway database maintenance. If a pathway is poorly understood or indeed poorly maintained then the number of gene interaction edges listed in its schema, and thus the pool of genes used, will be small and unrepresentative.

As a result I have implemented a second variation of the Pathway Entropy method: 'all' (PE-A) which takes for w_i all weighted edges between pathway all listed gene members, regardless of whether they are annotated with interaction information. This method may thus more reliably detect pathways whose genes are highly co-expressed outside of a pathway's established topology of interactions, particularly with regard to pathways whose functional links are less well known or understood. However as a result of not discriminating in its use of network edges it may increase the likelihood of false positives arising from co-expressed gene populations whose interactions are not known that are shared incidentally across multiple pathways.

The difference between PE-A and PE-T will thus vary from pathway to pathway dependent on pathway annotation. If a pathway is well understood and all genes within it are annotated to its topology of interactions then there will be no difference between the methods. The importance of the distinction arises in less well understood pathways which can contain tens of genes that are known to be associated with the pathway, but whose function within it is poorly understood. For these pathways PE-T therefore assesses pathway involvement conservatively using the sub-population of genes whose role in the pathway is known, whereas PE-A will additionally include all genes whose role is not well known.

3.2.2.3 *Demonstration of Entropy Calculation Using Experimental Data*

In this section I will demonstrate the results achieved for one of the gene co-expression networks I have generated. The network was generated using the OSm-MA dataset described in Section 1.4.1; this particular network was used in the sensitivity investigation in Section 3.3.3 and the methodology for its creation can be seen there. I have chosen this network as exemplar for the entropy calculation as it contains an artificially strengthened representation of the KEGG 'Long-term potentiation' (LTP) pathway and thus makes a good demonstration for the low-entropy scenario resulting from strong pathway representation.

The network, described in Table 3.1, is comprised of 16 clustered sub-networks, with a total of 17420 gene nodes. The number of genes and edges belonging to the

KEGG LTP pathway are listed for each cluster, additionally the sum of the Pearson's correlation coeffecients that weight the gene edges is included for all edges between between pathway members within a cluster. This correlation was calculated in a pairwise manner by WGCNA during network construction. The KEGG LTP pathway contained 68 gene members at the time of this analysis. Table 3.2 shows the cluster-wise probability, $P(V = i)$, of the pathway's association; the normalisations I have introduced, for pathway size and cluster size, are also demonstrated at this level both individually and combined. Lastly we can see the final pathway entropy for LTP in this network listed in the final row. The entropy methodology used for these calculations here is my PE-T method using the edge weights.

We can see that pathway size normalisation has the biggest effect on the entropy. The comparatively minor effect of the cluster size normalisation is the result of the connectivity for the largest co-expressed group of LTP genes representing but a small proportion of the overall number of edges in cluster 9 (see Table 3.1). That the final entropy scores are close to zero demonstrates that there is a strong well-ordered representation of the LTP pathway in this clustered network, which we can of course confirm as we have artificially strengthened its members' co-expression. Thus the Pathway Entropy method can be seen to function as intended in this scenario.

Cluster	Size(Genes)	Size(Edges)	Pathway Genes	Pathway Gene Edges	Pathway Correlation
1	758	49228	0	0	1
2	3024	615189	8	10	5.97
3	586	16473	4	2	0.43
4	712	23774	1	0	1
5	383	12033	1	0	1
6	3197	563241	7	7	2.15
7	5433	2470839	8	8	2.30
8	1102	66844	2	0	1
9	459	38670	34	1089	1089
10	596	15883	1	0	1
11	486	33837	1	0	1
12	138	1103	0	0	1
13	44	120	0	0	1
14	240	2942	1	0	1
15	62	247	0	0	1
16	165	1463	0	0	1

Table 3.1: **Example Cluster Summary**.
This table details the key descriptive features of each cluster. 'Pathway Genes' is the number of LTP genes in the pathway whereas 'Pathway Gene Edges' is the number of edges between these genes in the network; these do not necessarily represent extant functional relationships in the pathway. 'Pathway Correlation' describes the sum of the Pearson correlation values weighting edges that connect pathway genes, this was calculated in a pairwise manner for all genes in the network by WGCNA during network construction.

Cluster	$P(V = i)$	Pathway Normalised	Cluster Normalised	Pathway & Cluster Normalised
1	0	0	0	0
2	0.0281	0.0248	0.0281	0.0248
3	0.0031	0.0029	0.0031	0.0029
4	0	0	0	0
5	0	0	0	0
6	0.0121	0.0108	0.0121	0.0108
7	0.0128	0.0113	0.0128	0.0113
8	0	0	0	0
9	0.0195	0.0097	0.0189	0.0095
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
Entropy:	0.0756	0.0595	0.0750	0.0593

Table 3.2: **Cluster-wise Entropy Summary**.
This table summarises calculation of the entropy metric from the cluster-wise $P(V = i)$ and details the effects of the normalisation operations. The pathway entropy for each normalisation is listed in the final row. An exact '0' indicates that no result is possible as a consequence of the cluster containing no pathway edges.

3.2.2.4 *Calculation of Significance for Pathway Entropy*

In order to derive a significance value for the enrichment of a given pathway, the entropy value in question is tested against a null distribution of entropy values generated through permutation testing. For the DiNA method the authors simply took an agglomeration of all entropy values for all pathways to form this test distribution, however for the Pathway Entropy method I have deemed it more appropriate to test a pathway's entropy against a test distribution of values from that pathway. This is because pathways vary greatly in both size and connectivity which will necessarily impact and constrain the possible entropy values; if a pathway contains only a handful of genes, the limited number of connections between these members will constrain the results from the probability function in Equation 3.4. In turn this will limit the possible range of entropy values rendering it inappropriate to test for signif-

icance using an agglomerated distribution containing entropy values impossible to calculate for such a pathway.

Permutation testing to generate these test distributions is carried out by randomly shuffling gene labels in order to preserve the underlying topology of the network we are working with. This random shuffling is repeated over a sufficient number of iterations in order to produce a null distribution for testing. As a result test distributions are better fitted to the feature space of the data we are analysing.

An initial investigation into the shape of the test distributions, discussed further in results Section 3.3.2, demonstrated that they are not uniformly normal and as such non-parametric testing is required. A standard permutation test was employed, the detail of and justification for this is laid out in the context of this initial investigation in Section 3.3.2.

3.2.3 *The Pathway Entropy Pipeline*

In this section I will overview the methods and procedures that together form the computational pipeline that carries out the pathway entropy analysis. This will include detail and discussion of key parameters for relevant methods.

3.2.3.1 *Overview*

The pipeline consists of a sequential series of methods arranged and executed, with the exception of some preprocessing, entirely within the confines of the R programming language. Key methods such as network construction have been carried out utilising existing bioinformatics tools in R libraries and are appropriately detailed and credited in the following sections. The pipeline as a whole and the new pathway entropy calculation represent my novel contributions.

3.2.3.2 *Initial RNA-Seq Analysis & Pre-processing*

The first requisite step is to pre-process the RNA-Seq data to be analysed. The network construction requires the data to be in the format of count data, such as that produced by the `featureCounts` tool. The pipeline takes the count data and preprocesses it in accordance to guidance from the authors of the network construction tool WGCNA; genes which have counts less than or equal to 10 for 90% of the samples are removed due to insufficient representation and a variance stabilising transformation from the DESeq2 package is then applied to the data in order to normalise for the presence of batch effects. After this the data is supplied to WGCNA for network construction.

3.2.3.3 *WGCNA*

Weighted Gene Co-expression Network Analysis (WGCNA) is a well established tool and R package for the construction of weighted gene co-expression networks that has been widely used by bioinformaticians, as discussed in Section 1.3.1.1. It takes a set of experimental microarray or RNA-Seq data and applies a correlation method to the counts data in a pairwise method between genes, resulting in an adjacency matrix for the network. WGCNA utilises a 'soft thresholding' approach to network creation where it does not remove edges between genes that are below a certain threshold, instead it raises the absolute value of the correlation metrics by a 'soft thresholding power' to emphasise high correlation over low correlation. How this is chosen is described briefly in the following section.

WGCNA then clusters the adjacency matrix using hierarchical clustering to create a series of clusters or 'modules': sub-networks of the overall gene co-expression network that share a similar level of co-expression across the samples in the experimental data. These modules represent the final product of WGCNA and can then be used for downstream analysis. There are several key parameters in the network construction process that will be described further in Section 3.2.3.4.

Soft Thresholding

WGCNA's 'soft thresholding power' is a means by which it ensures scale independence in the gene co-expression networks it constructs. Scale independence (sometimes referred to as 'scale-free') is a feature of many networks (Barabási, 2009) that describes the fact that the overall connectivity of a network's nodes $p(k)$ follows a power law distribution, e.g. $p(k) \sim k^{-\gamma}$ (Zhao et al., 2010). Scale independent networks are extremely heterogeneous, their topology being dominated by a few highly connected nodes or hubs, and display a surprising degree of tolerance against errors (Zhang and Horvath, 2005). Scale independence has long been shown to be a characteristic of biological networks (Barabási and Albert, 1999), including gene co-expression networks (Barabási, 2009; Van Noort et al., 2004; Aggarwal et al., 2006), and therefore its presence in gene co-expression networks is important in order to be representative of the underlying biology.

The 'soft thresholding power' is not determined automatically by the WGCNA package. A function `pickSoftThreshold()` is run by the user, taking the counts data as input. This function returns a list of metrics (the scale free fit R^2, the slope of the power law function $slope$ and mean connectivity k) relating the the scale independence of the data for a series of power values. The user must then choose the value that best fits the criteria either of their choosing or as advised by WGCNA: $R^2 > 0.8$,

a value of slope approximating -1 and a high mean connectivity or k. These advised criteria are based on the criterion to approximate a scale independent topology and roughly present a trade off between scale-free topology and mean connectivity.

For all datasets to which I have applied the Pathway Entropy pipeline, I have picked the most appropriate power as defined by these guidelines.

3.2.3.4 *Network Construction Key Variables*

This section contains the description and details of the key parameters for the construction and clustering of gene co-expression networks in R using the WGCNA package. The parameter choices for this project are also detailed and explained.

Correlation Method

Several correlation methods are available for use in measuring the co-expression of genes. Whilst Pearson's correlation coefficient is the default in the WGCNA package, Spearman's rank correlation coefficient and biweight mid-correlation are also included as alternatives.

For the purposes of this project I have opted to use Pearson's correlation as it is the default and is frequently used when WGCNA has been applied in the literature (Luo et al., 2018; Bailey et al., 2016; Liu et al., 2017b).

Minimum Cluster Size

For the clustering of the network, WGCNA takes a parameter representing the minimum size a cluster can take, this is represented as the number of network nodes, thus genes, and does not measure edge connectivity. The default is 30 genes as the authors emphasise that bigger modules are more informative.

As this parameter will have a large effect on the calculation of pathway entropy, as Equation 3.4 is run for each cluster, I have run an investigation into the effect of decreasing the minimum cluster size parameter to its minimum value, 10, on the basis that this may produce more specific and functionally distinct clusters. This figure was reached through empirical trial as the number that, below which, too many clusters are generated using OSm-MA and WGCNA error quits. The results of this investigation can be seen in Section 3.3.5.1.

'deepSplit'

The parameter 'deepSplit' is a poorly documented sensitivity parameter involved in network construction in WGCNA. It can take an integer value in the range of $0 \leqslant n \leqslant 4$, with a default of 2, and is function is described in the WGCNA manual (Langfelder and Horvath, 2018) as follows: "Provides a simplified control over how sensitive module detection should be to module splitting, with 0 least and 4 most sensitive.".

It is described in more detail in another document by some of the same authors (Langfelder et al., 2007) where it is detailed to affect minimum gap and the maximum core scatter when applying the packages' dynamic tree cut function to the dendrogram produced by hierarchical clustering of the network, thereby affecting the sensitivity of the hierarchical clustering implementation when clustering the network.

As this parameter is a WGCNA key variable for creating the modules produced through its clustering method, I tested the increase of deepSplit to 4 to see, similarly to minimum cluster size, if more functionally distinct clusters may result. The results of this testing can be seen in Section 3.3.5.2.

Output Edge Weight Threshold

Whilst the WGCNA network creation method uses soft-thresholding to remove the need to exclude any potential connections between genes, no matter how small their correlated co-expression, it is impractical to save these networks for future analysis due to the high computational storage requirements of storing even a compressed all gene vs all gene edgelists for a network with a 17,000 x 17,000 adjacency list. This poses a particular problem for the generation of the null distributions for permutation testing for example, which requires the generation and storage of several thousand networks.

This is also an issue tacitly acknowledged by WGCNA whose network export functions include a default hard threshold requiring a minimum correlation of 0.5 for an edge to be written to the edge list. In order to preserve as much co-expression information as possible I have used a lower threshold of 0.2. This figure was chosen through empirical testing which demonstrated lower thresholds to increase file sizes by orders of magnitude. This threshold produces a compressed edge list of approximately 75MB in size using the OSm-MA data set detailed in Section 1.4.1.

3.2.3.5 *Entropy Method & Key Variables*

In this section I will briefly outline the key variables and methodology choices that can be made for the pathway entropy portion of the pipeline.

Use of Edge Weights

The use of edge weights, which are the correlation of co-expression between two given genes, to inform pathway enrichment is optional in the Pathway Entropy methodology. Whilst it is one of the key features I have introduced, on the basis that including this information can only make resultant enrichments more representative of the source data, I have added in the option not to use it as a test case for comparison. The use or not of edge weights will be labelled in the results section through reference to whether a Pathway Entropy method is 'weighted' or 'unweighted'.

Normalisation Procedures

The user can specify whether they would like to normalise for pathway size, size of clusters containing pathway members or both.

The pathway size normalisation is carried out on the basis of gene member representation; a cluster which contains a larger number of pathway gene members will see a reduction in entropy as there is less disorder in the pathway's representation, whereas a cluster with few pathway members will receive a less favourable reduction.

The cluster size normalisation is carried out in similar fashion; however it is done on the basis of edges rather than gene nodes. A cluster or module whose connectivity is entirely between pathway gene members will receive a privileged decrease in entropy in recognition that the co-expression of the pathway genes is strong and unique enough to be clustered separately from the rest of the network. Conversely if a pathway's connectivity is represented in a large cluster of unrelated genes and connectivity, the normalisation will be marginal.

Limit to Established Functional Knowledge

As previously described in Section 3.2.2.2, there are two variants to my implementation of the Pathway Entropy methodology: PE-T and PE-A. The former places emphasis on established functional knowledge, using only the subset of gene members present in a pathway's functional topology, whilst the latter places no such limits, using all available co-expressed relationships between all pathway gene members regardless of their presence in a pathway's functional topology. The choice of either variant is thus a key one in the pipeline and one analogous to prioritising precision or recall; PE-T, drawing input only from genes that have edges which represent established biological knowledge, will be less likely to produce false positives, PE-A on the other hand will be more likely to avoid false negatives through taking all available

information between all pathway gene members into account. The PE-A method as mentioned previously will be inherently more resilient to poor pathway maintenance or limited functional knowledge as representation in a pathway's functional topology is not required for gene members to be used.

3.2.3.6 *Permutation Testing & Significance Calculation*

The final portion of the Pathway Entropy pipeline is to attribute significance to pathways based upon their entropy value. This is carried out through testing each pathway's entropy value against a pathway specific null distribution generated from the same dataset as described in Section 3.2.2.4.

The null distributions are generated by constructing the gene co-expression network for the dataset and then shuffling the gene labels at random and saving the randomised network to file. This is carried out for as many iterations as the user wishes after which the entropy values for each pathway are extracted to form the pathway specific null distributions for significance testing. The null distributions are saved to file once extracted and so precluding the need for regeneration for any repeat or subsequent analysis of the same dataset.

The advantage of permutation testing is that it is fast, after initial data generation, and preserves the topology of the network thus meaning the resultant null distributions are based on permutations of possible entropy values for a network of such a structure. This is also known as the principal of 'exchangeability', where any configuration of data points is just as likely as the original.

3.2.3.7 *Source Information for Network Construction*

The KEGG pathway database is primarily composed of pathways of protein coding genes, with a small minority of pathways including information from short non-coding RNA genes (e.g KEGG pathway 05206 - 'MicroRNAs in cancer') or pseudogenes. Whilst the presence of non-coding genes will only directly affect the analysis of the PE methods in a negative manner through cluster size normalisation, as they otherwise use only information relating to the genes of the pathway they are analysing, they will have a greater impact on the analysis of comparative non-entropy based methods. For hypergeometric methods the presence of non-coding genes, whose presence will not contribute to the enrichment for the majority of pathways, will inflate the test statistic and can result in lower p-values for enriched pathways.

Whilst the protein coding networks are arguably the most relevant to the pathway enrichment, the removal of gene expression information for non-coding genes will necessarily change the the structure of the networks and reduce their overall information content. The inclusion of non-coding information in these network may

also be important in the future as the incorporation of non-coding genes into pathway databases will likely increase with our knowledge of their roles in biological processes.

In order to address both points I will, in the comparative results sections of this chapter, present analysis from networks constructed using both all available information and also networks restricted to only information from protein coding genes.

3.3 RESULTS

3.3.1 *Entropy Pathway Distributions*

In this section, I will demonstrate the distributions of entropy generated both on the level of individual pathways in situations where there is poor as well as rich representation in the data, but also generalised over all pathways. This will demonstrate the range of values generated by the Pathway Entropy pipeline so that appropriate methods for significance testing can be determined.

On the level of individual biological pathways, there are two key factors that will affect the calculated entropy value: pathway size and the proportion of pathway members represented within the source data.

Larger biological pathways will necessarily incorporate a wider range of biological processes and functions which increases the chance that the genes involved in specific sections may be clustered together separately from the rest of the genes in the pathway despite sharing a greater overall function. Indeed some KEGG pathways like KEGG 5010 ("Alzheimer's Disease") include significant representation of other pathways which are in fact constituent components within them, such as KEGG 190 ("Oxidative Phosphorylation"). In cases such as these, depending on coexpression levels within the network, we may expect genes in this pathway to split across multiple clusters if there is a greater coexpression between genes shared with the Oxidative Phosphorylation pathway. Where pathway members are split across multiple clusters we will observe a notable change in entropy value, as the overall value will now be derived from the sum of constituent values provided by each cluster containing pathway genes. If coexpression, and thus correlation, is high between pathway members then the overall entropy results may still be low as pathway involvement is more certain. Similarly in situations where we have only marginal representations of stray members in other clusters, the small values contributed by these should not significantly affect the overall entropy.

Representation of a biological pathway within the source data, that is the expression of the pathway's gene members, is a key factor in entropy calculation for two key reasons: normalisation and permutation testing. As we normalise for the proportion of pathway membership present in the clustered data, poor pathway membership will result in normalisation to a higher entropy value to reflect the diminished likelihood of the pathway's involvement. If a pathway's members are not well represented in the source data, however, the null distribution derived from the permutation bootstraps will be sparser and as a result, the ability to achieve high significance when testing against it thus impaired. Indeed if the null distribution is too sparse for a

pathway, significance will not be calculated. However I will explain this further in the following section.

The effect of poor pathway representation where members are split over multiple clusters can be seen clearly in Figure 3.2A & B. This figure describes the combined null distribution over all represented pathways tested from a base of 321 individual KEGG pathways. The tri-modal nature of this distribution can be explained quite neatly by the number of clusters the pathway members are partitioned into; the first peak of lowest entropy represents permutation bootstrap iterations where pathway members are gathered into a single cluster, similarly the second peak represents permutation bootstrap iterations where members are partitioned between two clusters and lastly the final peak representing iterations where members are split between three or greater clusters. This tri-modal distribution is seen across all pathways that are poorly represented in the data, or whose membership is small, whose entropy values will be most affected by the partitioning of gene membership. The peaks are in part also a result of the discrete number of possible entropy values which is visible particularly in situations of less data such as poorly represented pathways. With a sufficiently large dataset and pathway it may be possible to discern additional peaks, as this would allow more for more connectivity and therefore a wider range of possible values, however using this dataset with the KEGG pathway database there is not a wide enough range for pathways represented in three clusters to be distinct from those represented in more than three. Whilst a tri-modal distribution is not as ideal for statistical testing as a normal distribution, that it is so clearly a result of membership representation in the clustering is unavoidable.

For individual pathways, of moderate size, that are well represented in the data and that do not contain nested pathways we tend to see an approximation of a normal distribution of entropy values. This can be seen in Figure 3.2C & D for the KEGG 'Glycolysis / Gluconeogenesis' (KEGG accession 10) pathway where 52 out of 66 pathway gene members, or 78.8%, are represented. Indeed for approximately a third of the 321 KEGG pathways tested, a normal distribution was observed. The lack of tri-modal distribution in these cases is due to the larger number of pathway genes meaning that even in the case of partitioning, there is such a volume of data as to provide a wider possible range of entropy values. For visual comparison a poorly represented individual pathway can be seen in Figure 3.2E & F where only 36 out of 1100 pathway gene members, or 3.3%, are represented in the network data.

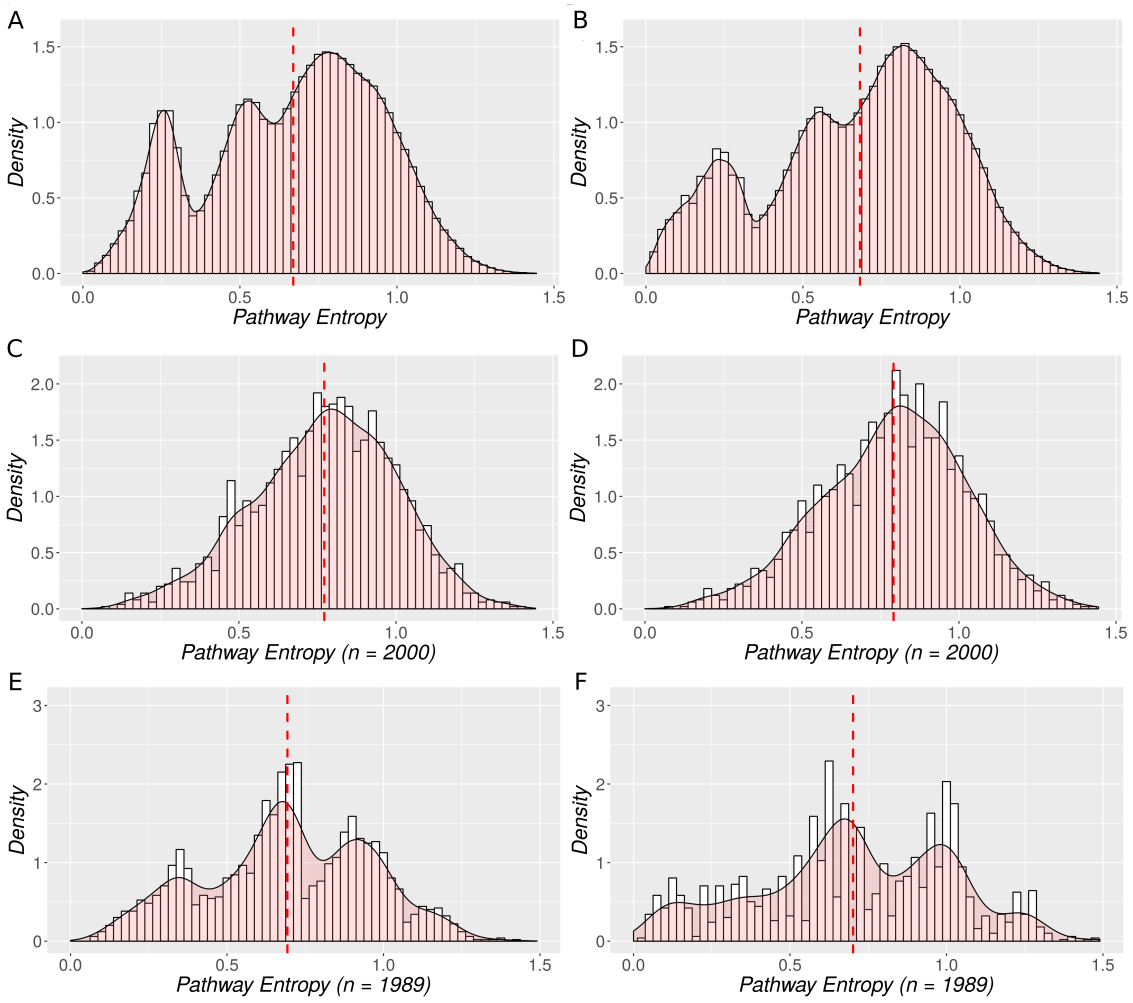


Figure 3.2: **Entropy Permutation Distributions.** (A) & (B) Show the combined entropy distributions of all represented pathways for the Weighted and Unweighted Topology entropy methods respectively, the former utilises the edge weights between nodes of the network (the correlation values themselves) whilst the latter utilises a simple count of edges between pathway member. (C) & (D) Show the entropy distribution for the well represented KEGG 'Glycolysis / Gluconeogenesis' (10) pathway, which tends toward a normal distribution, as generated again by the Weighted and Unweighted Topology entropy methods respectively, whilst (E) & (F) show the same for the poorly represented KEGG 'Olfactory transduction' Pathway (4740). The methods used to generate all these graphs have been normalised for pathway size and connectivity. 'n', on the x-axis label of figures C-F, refers to the number of permutation bootstrap iterations that produced an entropy value for that pathway.

3.3.2 *Significance Calculation*

As described in Section 3.2.2.4, the Pathway Entropy method developed here uses iterative permutations of gene labels, in a bootstrapping fashion, to create null distributions for the permutation testing of entropy significance. Each permutation is generated from the same source data that is to be analysed by shuffling the gene labels and calculating the entropy for each pathway represented in the data. These entropy values are collated into the pathway specific null distributions that can be seen in Figure 3.2C-F. I have used 2,000 permutations to generate these distributions as this was enough for the generated p-values to be stable.

As the pathway distributions vary considerably based on membership representation in the source data, and indeed to account for variation in membership between pathways, it was considered most appropriate to perform any significance tests for a pathway's entropy values against that pathway's own null distribution. This maximises the use of available information in order to create more reliable significance values, a marked difference between this method and the DiNA method developed in (Gambardella et al., 2013), which simply uses an overall null distribution from all the pathways they are testing.

Given that only the most well-represented pathways in the data will approximate a normal distribution, parametric statistical tests could not be used. Non-parametric statistical tests were investigated and considered, however many commonly used tests (e.g. Wilcoxon signed-rank) simply test the distance of the value from the population average and thus are not appropriate for testing single values (e.g a pathway entropy) against distributions with a wide range of values. Thus I chose a standard permutation test in order to determine significance as it allows direct comparison with DiNA and also is appropriate for non-parametric data.

To test an entropy value e against a null distribution N with members i and of length l, we use the following equation to calculate an empirical p-value:

$$p = \frac{N_i < e}{l} \tag{3.5}$$

This method has two key disadvantages: firstly, the accuracy is determined by the number of permutations carried out, the disadvantage here is that the computational cost increases with each permutation both in terms of calculation and in terms of data storage, thus a balance must be struck between precision and computational practicality. Secondly, that the significance value achieved has a smaller feature space than is usual for statistical tests of significance. Regarding the first point; 2000 per-

mutations were used to generate the null distributions as this was observed to be sufficient precision for both generated p-values to rarely hit the lower value limit and to not present a data storage issue. The second point means that it is harder to directly compare significance between the Pathway Entropy method and existing alternatives for pathway analysis. However given that this tool is to be used primarily as an indicator of pathway involvement, a smaller significance feature space can be seen as an advantage.

It is possible that the number of permutations could constrain the use of the method as an evaluator of network construction methodology as the significance cap could prevent comparative evaluation if hit by multiple pathways when comparing methods. However, as this has not been the case when testing the Pathway Entropy method on real experimental data I do not anticipate this problem hindering our investigation. If it were, increasing the number of permutations should be sufficient to relieve this issue.

I will now compare this methodology to the most similar existing method. In (Gambardella et al., 2013) the authors test for significance in a similar manner; they test an entropy value also using permutation testing, against a null distribution generated by the same gene shuffling permutation procedure of 10,000 repetitions, combining values for all pathways. There is one key differences here: the use of a combined null distribution. The use of a non-pathway specific null distribution for significance testing would be an inappropriate choice for my methodology as it would not take advantage of pathway specific information, also potentially violating the idea of exchangeability in permutation testing as not all entropy values could be generated for all pathways, and would mean the use of a tri-modal null distribution for all pathways thus improperly testing pathways whose distribution differs. It is important to note here that as (Gambardella et al., 2013) threshold the pathways used, requiring 80% gene representation, their distribution may look markedly different and whilst it can certainly be argued that such pruning of pathways in advance of analysis may be expedient, for a laboratory biologist assessing the results of pathway analysis, it can be useful to have less well represented pathways included to allow full investigation and interpretation of the data.

3.3.3 *Sensitivity*

We next investigated the sensitivity of the Pathway Entropy method to expression changes in the source data. This section will describe the methodology used to artificially simulate expression change *in silico*, via the addition of noise to the OSm-MA dataset (described in Section 1.4.1), the results achieved and will examine whether

the method is sensitive enough to expression change to be suitably applied as both an enrichment and evaluative tool.

A common base methodology was used for these investigations. Firstly, gene counts for the gene members of a chosen pathway were set to be equal. Secondly, for a three condition RNA-Seq experiment with three replicates in each condition, the counts for pathway gene members were fixed at 1,000 for conditions one and three and then increased by a chosen fold change for condition two. This was applied uniformly across replicates. Thirdly, any pathway members not present in the source data were added in. Default or recommended values were used for network construction, the networks were constructed using only information from protein coding genes.

3.3.3.1 *Noise Methodology*

It is important that the noise modelled be biologically realistic. Thus it seemed appropriate to artificially perturb a pathway within an actual RNA-Seq dataset so that all but the deliberately manipulated gene counts are biologically realistic by the fact of them being unaltered. Taking this as a base consideration for a noise methodology two different approaches were considered.

The first approach was to apply random noise in a series of increasing noise ranges, e.g. 0 - 500, 500 - 1,000. A random value between the upper and lower bound would be randomly added to or subtracted from the count for a particular gene member for an individual sample replicate, with the process repeating for every sample replicate. Whilst this method produced more realistic noise, the choice of noise bands became a non-trivial issue. If the bands were too large then the effect would be so different between replicates as to be incomparable; however if they were small then this substantially increased the number of tests to run. With respect to the latter, this proved a costly effort both in processing time and computational power. As such an alternative approach was investigated.

The second approach was to take the uniform count perturbations and instead of applying the noise directly to the counts, introduce it through shuffling gene labels. After the adjacency matrix of the network had been constructed, but before the clustering stage, a specified percentage of the pathway gene membership would be randomly shuffled with other gene labels. A control was added to prevent the edge scenario in which pathway members swapped labels with one another and thus introduce no noise. For this approach we had the advantage of being able to reliably and comparably distort the representation of the pathway in the data. Six levels of noise were chosen in order to observe the distortion at reasonably spaced intervals in addition to a non-distorted control: 0%, 5%, 10%, 25%, 50%, 75%, 95%. The largest

noise percentage was chosen to be close to but not exactly 100% as at this level there would be no longer a pathway signal to detect.

3.3.3.2 *Results of Noise Application*

The noise methodology described above was applied to two different KEGG pathways that are well represented in the data: 'Glycolysis / Gluconeogenesis' (10) with 78.8% membership representation and 'Long-term potentiation' (4720) with 91.1% membership representation.

This section will detail the results of these two pathway perturbations separately before making some comparative comments.

Pathway Perturbation of KEGG pathway 'Glycolysis / Gluconeogenesis'

I will present first the results of the Pathway Entropy methods before comparing these as far as possible with the DiNA methodology. I have implemented the original DiNA as detailed in (Gambardella et al., 2013), however I have also implemented a second version that restricts edges counted to only those explicitly listed in the KEGG pathway database (edges thus representing biological function) in order to allow for comparison with my PE-T method. Significance, as q-values (Benjamini-Hochberg multiple testing correction) were calculated from the raw p-values for each noise level (0%, 5%, 10%, 25%, 50%, 75% and 95%) and subsequently repeated for 5 different fold change magnitudes: 0.5, 1, 2, 5 and 10. This procedure was then repeated to obtain 3 values for each noise level in order to calculate the standard error of each data point.

The results for KEGG pathway 'Glycolysis / Gluconeogenesis' (10), hence referred to as 'Glycolysis', can be seen in Figure 3.3. Each method has its own colour and data points are round for the unweighted, edge counting methods and triangular for the weighted, edge weight summation methods. From this graph, we can see a similar performance across all entropy based methods, particularly between the PE-T and PE-A approaches, with a decrease in enriched q-value with an increase in noise. For the DiNA method we can see that this decrease happens more abruptly, at either 25% or 50% noise dependent on fold change, with the pathway entropy approaches' q-values decreasing much more gradually with noise. However performance between the entropy methods is so close as to only be significantly different ($p < 0.05$, Wilxox signed-rank test) between pathway entropy approaches and DiNA approaches at a noise level of 0 for each fold change ($p = 0.04$), and also for 5% noise on fold change 0.5 ($p = 0.04$).

Although the DiNA approaches significantly outperform the pathway entropy methods at an instance of perfect correlation, such a signal with zero noise is not biolog-

ically realistic. It is thus important to note that the degree of enrichment achieved by the pathway entropy methods is not significantly different to that of the original methodology. Whilst the q-values produced by the pathway entropy methods are themselves not significantly different from the DiNA approaches at any fold change beyond 5% noise, it is worthy of note they have produced a significant enrichment robust to a higher level of noise for each fold change.

That both PE-T and PE-A methods perform so similarly here to be expected as the information content they are working with - perfect correlation based on exact co-expression - is very high even at relatively large levels of noise. However that even with such strong co-expression the perturbed pathway signal is lost at low fold change magnitudes as can be seen at the 95% noise level on the 0.5 fold change graph in Figure 3.3.

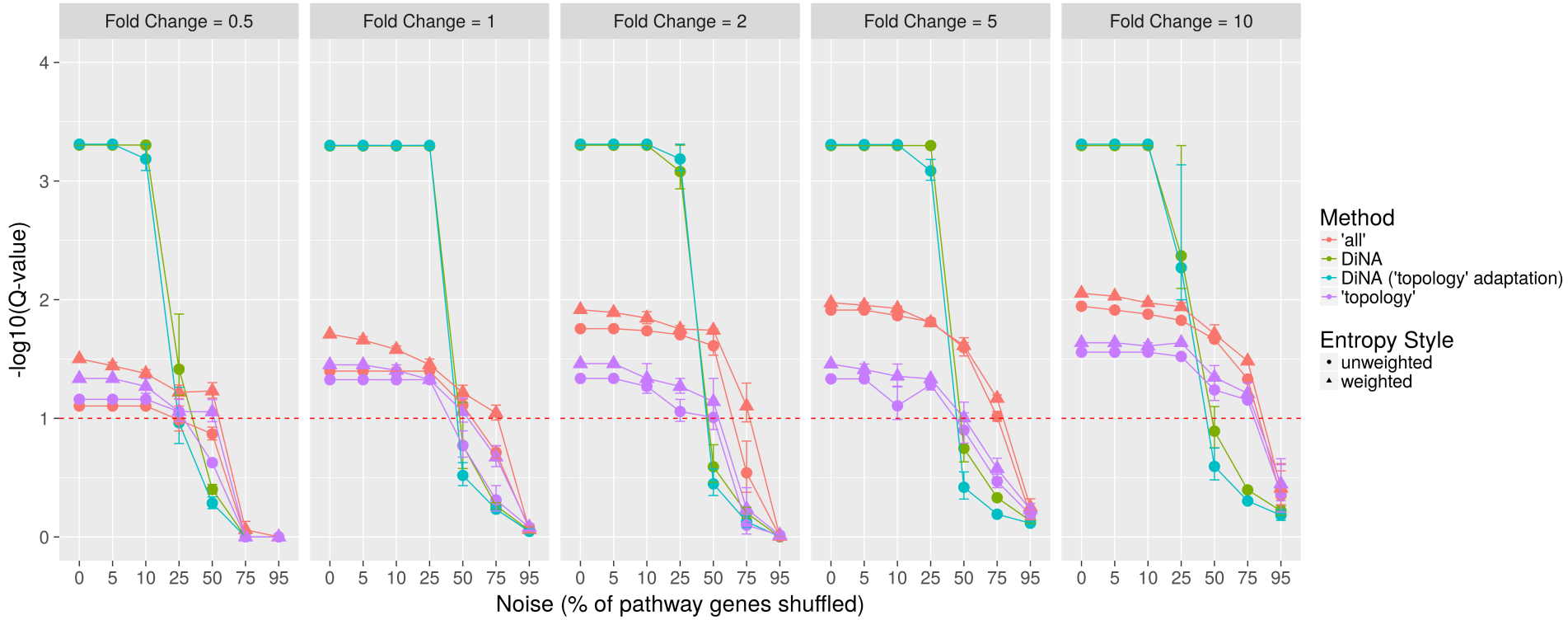


Figure 3.3: **-log10(q-value) (Benjamini-Hochberg) Significance for Application of Entropy Methods to Network with Perturbation of KEGG 'Glycolysis / Gluconeogenesis' Pathway.** This series of graphs displays the -log10(q-value) generated for the DiNA, both regular and 'topology adapted', PE-T and PE-A approaches. The weighted methods can be seen to produce marginally more significant enrichments than the unweighted methods for the Pathway Entropy methods, with the DiNA methods generating the lowest q-values for lower levels of noise. However the q-values produced by the two method groups are only significantly different ($p = 0.04$, Wilcoxon rank) when noise is absent or for 5% noise when the fold change is 0.5. The q-values were generated through the application of the Benjamini-Hochberg multiple testing correction. The error bars displayed show the standard error for each data point, taken from n=3 replicates. The dashed red line represents a typical q-value cut-off of 0.1, with negative log scaling values below this threshold are displayed above this line.

Pathway Perturbation of KEGG pathway 'Long-term potentiation'

The pathway perturbation investigation was repeated with another well represented KEGG pathway: 'Long-term potentiation' (LTP), KEGG accession 4720. The results can be seen in Figure 3.4.

For the LTP pathway we see a comparable performance to that for the Glycolysis. Indeed once again the only significant ($p < 0.05$, Wilxox signed-rank test) difference in q-value produced is for the better performance of the DiNA approaches for zero noise, however this degree of performance difference is consistent at $p = 0.04$. For a fold change of 0.5, there is no longer significant difference at the 5% noise level. Again, whilst the sensitivity of the entropy based methods is not significantly different in the presence of noise, the pathway entropy methods have produced a significant enrichment robust at a higher level of noise for each fold change.

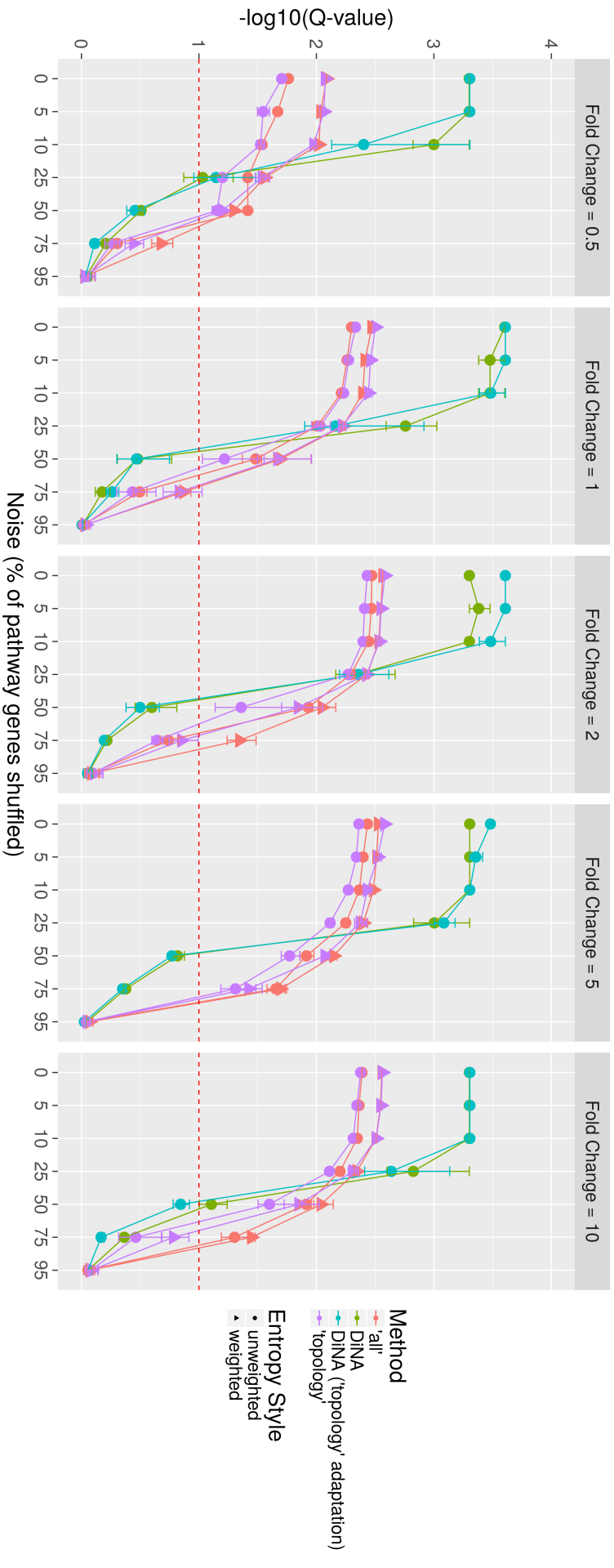


Figure 3.4: **-log10(q-value) (Benjamini-Hochberg) Significance for Application of Entropy Methods to Network with Perturbation of KEGG 'Long-term potentiation' Pathway.** This series of graphs displays the -log10(q-value) generated for each of the entropy based methods. Here too, the weighted methods can be seen to produce marginally more significant enrichments than the unweighted methods for the Pathway Entropy methods, with the DiNA methods generating the lowest q-values for lower levels of noise. The q-values produced here by the two method groups are only significantly different ($p = 0.04$, Wilcoxon rank) when noise is absent. The error bars displayed show the standard error for each data point; taken from n=3 replicates. The dashed red line represents a typical q-value cut-off of 0.1, with negative log scaling values below this threshold are displayed above this line.

3.3.4 *Comparative Performance Against Existing Entropy Methods*

Having described the base functionality and sensitivity of the Pathway Entropy methodology in the previous sections, I will now endeavour to compare it with existing methods. In section 3.3.4.1 I will compare the performance of the PE-T and PE-A methods with an implementation of the DiNA method created by (Gambardella et al., 2013) and a DiNA variant using the 'topology' methodology. In section 3.3.6 I will compare the Pathway Entropy methods against existing and long-standing gold standard hypergeometric methods for pathway enrichment using the KEGG pathway database: KEGGprofile and clusterProfiler.

These two sections will detail the same comparative investigations for both types of method. Whilst an investigation into sensitivity has already been conducted between entropy methods in the previous section, a comparative investigation will be carried out additionally for the hypergeometric evaluation.

3.3.4.1 *Between Entropy Methods*

In this section I will compare the new Pathway Entropy methods, PE-A & PE-T, both weighted and unweighted, with the DiNA method and the DiNA method adapted to the 'topology' methodology. This investigation will be carried out by applying these methods to a set of RNA-Seq data, looking at differences in entropy, pathway enrichment, enriched pathways ranked highly by one method but lowly by another and will also look at whether the size of a given pathway has any effect on the enrichment results derived. The dataset described in Section 1.4.1 was used for this investigation, networks were constructed using both all available and protein coding only expression information and are labelled as appropriate.

Entropy Variation Between Methods

The first comparison of the entropy methods is that of the entropy values they generate, before significance is calculated. An overview of the entropy values generated by the methods, whilst not necessarily indicative of their resultant significance, is useful for interpreting these downstream results. Figure 3.5 displays the entropy value generated by each entropy method: PE-A, PE-T, DiNA & the DiNA 'topology' adaptation with the weighted and unweighted variants of the former two displayed separately. The KEGG pathway identifiers for each pathway are listed along the x-axis. Pathways that are not represented in the data, and thus who have no entropy, are not shown. This data can also be seen as a violin plot in Figure 3.10A.

First note that the results are asymmetrical. The PE-A method and the DiNA method, which share a similar approach, return a full set of entropy values, whereas the methods of the 'topology' approach show results for only a subset. This is a result of functional topology not being universally annotated, or perhaps known, for all pathways in the KEGG database; as the 'topology' methods rely on counting only the subset of genes represented in these functional links, they cannot derive an entropy score for pathways that lack this information.

The log scaled entropy scores themselves vary over a large range from 0 to past 1.5. Looking through the heatmap there is a stronger representation of low -log10 entropy values for the DiNA and PE-A methods compared to PE-T, though less of a distinction between the weighted and unweighted Pathway Entropy methods. The former is almost certainly a result of the more stringent criteria in the PE-T methodology, whereas the latter is interesting. Whilst differences between weighted and unweighted can be observed, for pathway 5033 'Nicotine Addiction' for example, they are not common indicating that the additional information does not impart as much difference as we had expected. Whilst the differences between the weighted and unweighted variants are marginal in most cases, it is reassuring to see that using this additional information does indeed have an impact, even if minor, as expected.

PE-T, whilst returning less results than the PE-A method, appears to derive higher -log10 entropy values for many pathways, thus potentially highlighting situations where high representation of a pathway's core functions in the data is being obscured by the lack of overall membership representation or of high correlation between all members. This situation is not unexpected for pathways with larger gene membership as these will have a high number of potential links between gene members but a comparatively small functional topology. Differentially scored situations such as these will be investigated further in section 3.3.4.1.

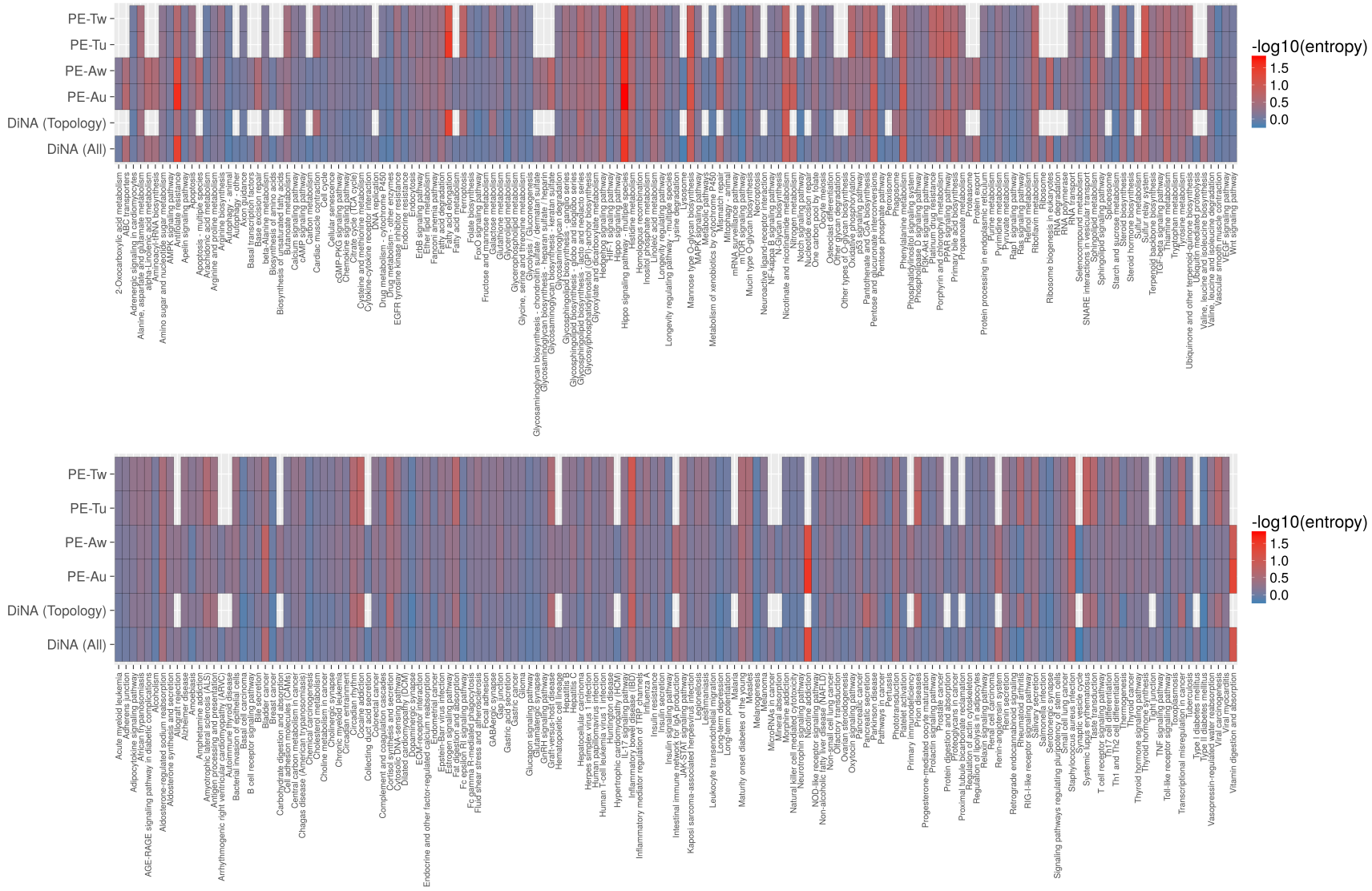


Figure 3.5: **Pathway-wise Entropy Values by Method.** This figure displays the entropy metric, scaled by -log10, generated by six different entropy based methodologies for each tested KEGG pathway; PE-A & PE-T, both weighted and unweighted, in addition to DiNA & DiNA 'topology' adaptation. Here we can see that all methods produce a full range of entropy values, whilst the entropy for many pathways can be seen to be of similar magnitude across all pathways, the are also distinct differences observable between the 'topology' based approaches and the PE-A and DiNA approaches. The x axis consists of KEGG pathway ids. Entropy has been visualised unbounded with a lower entropy indicating a higher certainty of pathway involvement in the dataset, 0 being certain. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

Pathway Enrichment

After applying the six different methods to the real dataset described in section 1.4.1, the pathway enrichment results were collated and plotted against one another by method as a heatmap to display the comparative significance of the KEGG pathways. Benjamini-Hochberg multiple testing correction (MTC) was applied to the p-values generated by the pathway enrichment.

Figures 3.6 and 3.7 show a comparative overview of pathway enrichment for both a network using all expression information (henceforth referred to as the 'all expression network' or 'AEN') and the network built from only protein coding gene co-expression (the 'protein coding network' or 'PCN'). For both we can clearly see that the majority of pathways have a poor q-value, which is not unexpected for a database of over 300 pathways. Where low q-values are observed in Pathway Entropy methods we see a similar performance between weighted and unweighted variants and often between PE-A and PE-T methodologies. For several pathways in the AEN, such as 'Non-alcoholic fatty liver disease', which contains the oxidative phosphorylation pathway as mentioned previously, and 'Circadian rhythm', which is an integral biological pathway, there is demonstrable cross-method consensus. In the PCN however there is less consensus, with only 'Hippo signaling pathway - multiple species', which covers cell survival and apoptosis, and 'Antifolate resistance', demonstrating a similarly lower q-value across all methods.

Once we limit this enrichment by the standard threshold for q-value significance ($q < 0.1$), as shown in Figure 3.8 for the AEN, we see a very different picture. In the AEN, the PE-T and the DiNA 'topology' adaptation produce no q-significant results, PE-A produces 8 and 12 q-significant pathways for weighted and unweighted respectively and DiNA produces 3 of which 2 of these are shared with PE-A. For the PCN it is only DiNA that produces q-significant enrichments: 2 out of the 3 enriched in the AEN.

Given the relative performance of the DiNA methods in the sensitivity investigation (Section 3.3.3) it is interesting to see that even though the DiNA method has the potential to be more robust after MTC due to the lower potential p-value, as discussed in Section 3.3.3.2, it performs comparatively poorly against the PE-A method in the AEN network in terms of q-significant pathway enrichment. That 2 of 3 of q-significant enrichments reported by DiNA survive the loss of information in the PCN yet none of PE-A's do highlights a sensitivity of the PE methods to information content in the network. Given that DiNA is less affected this would indicate that this sensitivity is a result of the additional normalisations in the PE methodology, of

which the cluster size normalisation may now be harsher with the reduction in the number of clusters.

In the AEN results, if we look at what pathways these methods are ascribing q-significance to, all three methods list 4392 ('Hippo signaling pathway - multiple species') and 4710 ('Circadian rhythm') as q-significant. The Hippo signaling pathway deals with, amongst other responses, a response to cellular stress stimuli and apoptosis related functions. The largest co-expressed group of pathway members represent exactly this portion of the pathway and thus so fits our experimental data perfectly as an appropriate pathway enrichment. The circadian rhythm pathway represents core biological function which we would expect to be represented in the data though not necessarily so strongly co-expressed. The only pathway that DiNA finds uniquely significant is the 'Antifolate resistance' pathway. Whilst the fact that antifolates deal with inflammation is broadly relevant to the inflammatory effects of oxidative stress and the pathway entry includes a related portion of the nF-kB signaling pathway whose role in cell-survival is relevant, there were no antifolate drugs used in the experimental procedure and so the significant enrichment of this pathway is not useful.

In AEN, the PE-A methods have a consensus on 8 q-significant pathways, in descending order of significance: 190 ('Oxidative phosphorylation'), 3008 ('Ribosome biogenesis in eukaryotes'), 4392 ('Hippo signaling pathway - multiple species'), 5010 ('Alzheimer's disease'), 4710 ('Circadian rhythm'), 4010 ('MAPK signaling pathway'), 5168 ('Herpes simplex infection') and 5012 ('Parkinson's disease'). The unweighted variant produces an additional 2: 601 ('Glycosphingolipid biosynthesis - lacto and neolacto series') and 4927 ('Cortisol synthesis and secretion'). Oxidative phosphorylation, as previously detailed, is eminently relevant to our dataset and so the presence of both Alzheimer's and Parkinson's disease pathways is a likely consequence of the representation of the oxidative phosphorylation pathway within these pathways.

Whilst PE-T did not produce any q-significant enrichments, Figure 3.7 shows that there is reasonable agreement between PE-A and PE-T. This is especially true for those pathways found q-significant by PE-A. Thus it would appear that it is PE-T's more stringent methodology here that is preventing enrichment.

Figure 3.6: **Pathway-wise q-values by Method for the AEN.** This figure shows the q-values generated by six different entropy based methodologies on the AEN for each tested KEGG pathway: PE-A, weighted and unweighted, PE-T, weighted and unweighted, in addition to DiNA & DiNA 'topology' adaptation. The PE-A approach can be seen to generate the widest range of q-values with only the DiNA approach enriching any pathways to a high -log10(q). The 'topology' based approaches can be seen to perform poorly at this level. The x axis consists of KEGG pathway ids. The q-values here have been visualised without upper bound. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

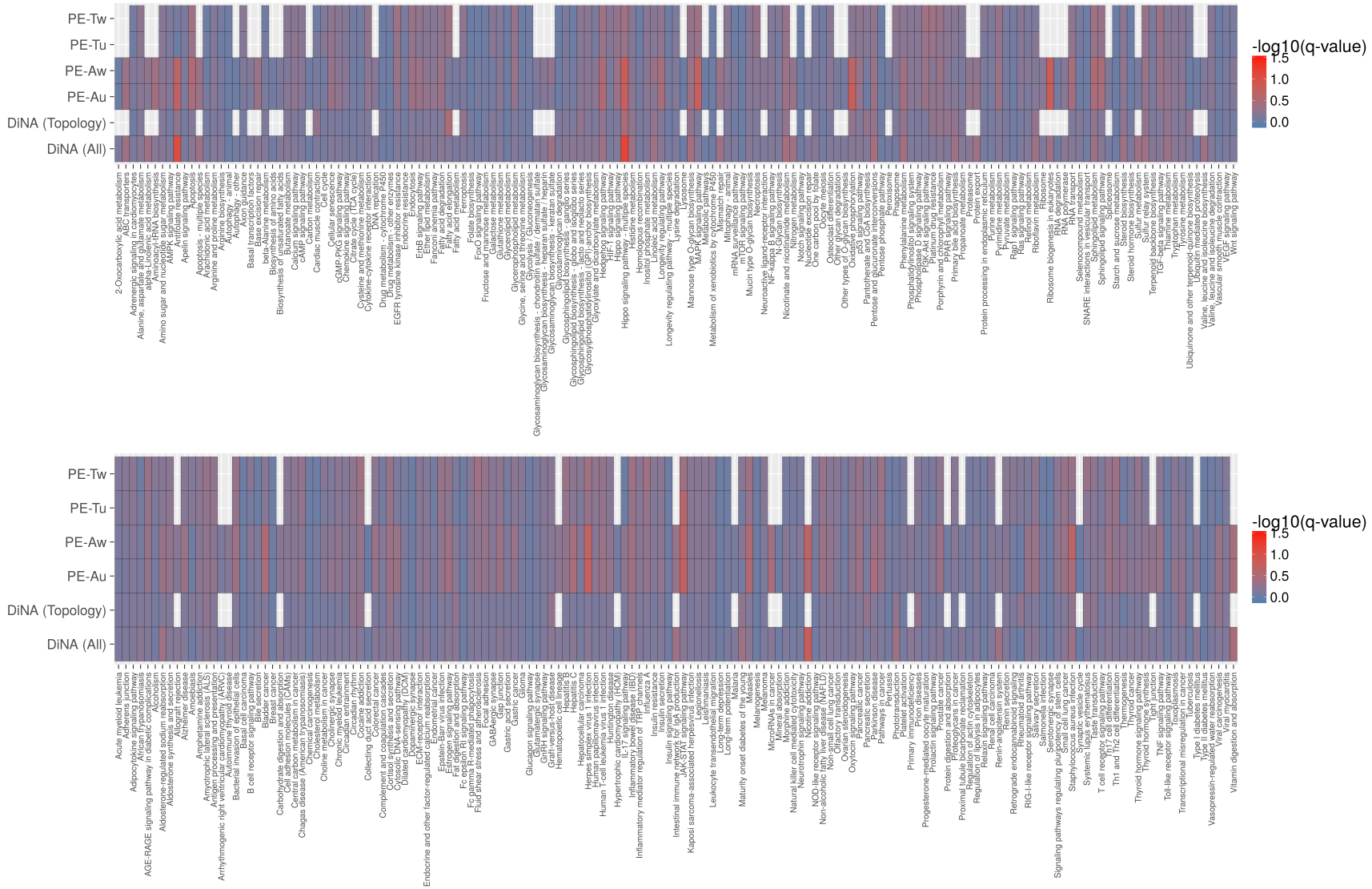


Figure 3.7: **Pathway-wise q-values by Method for the PCN.** This figure shows the q-values generated for the six different entropy based methodologies on the PCN for each tested KEGG pathway. We can see a noticeably worse performance for the PCN compared to the AEN in Figure 3.6, with few pathways garnering a high -log10(q) for any method, likely a result of the loss of network complexity. The x axis consists of KEGG pathway ids. The q-values here have been visualised without upper bound. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

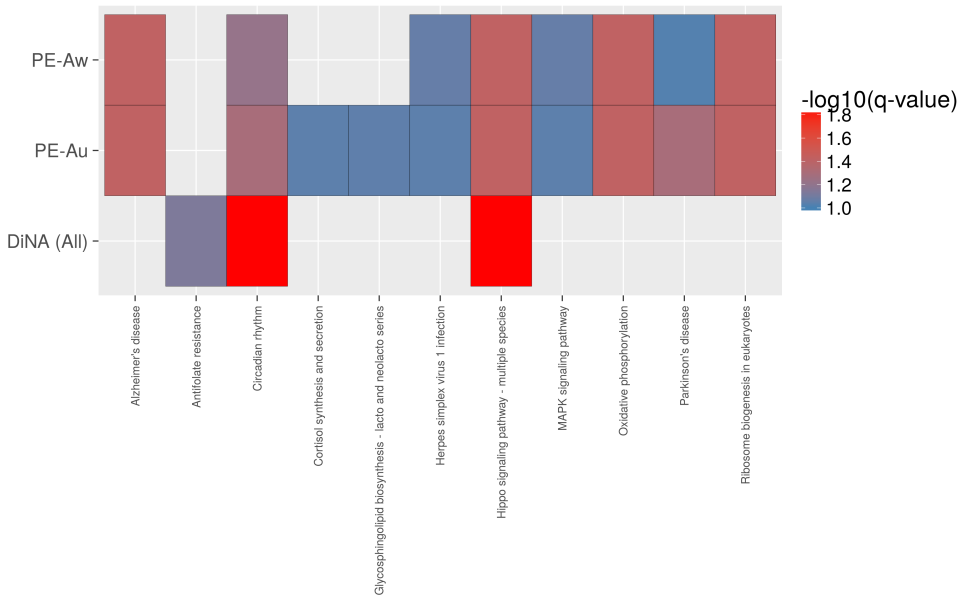


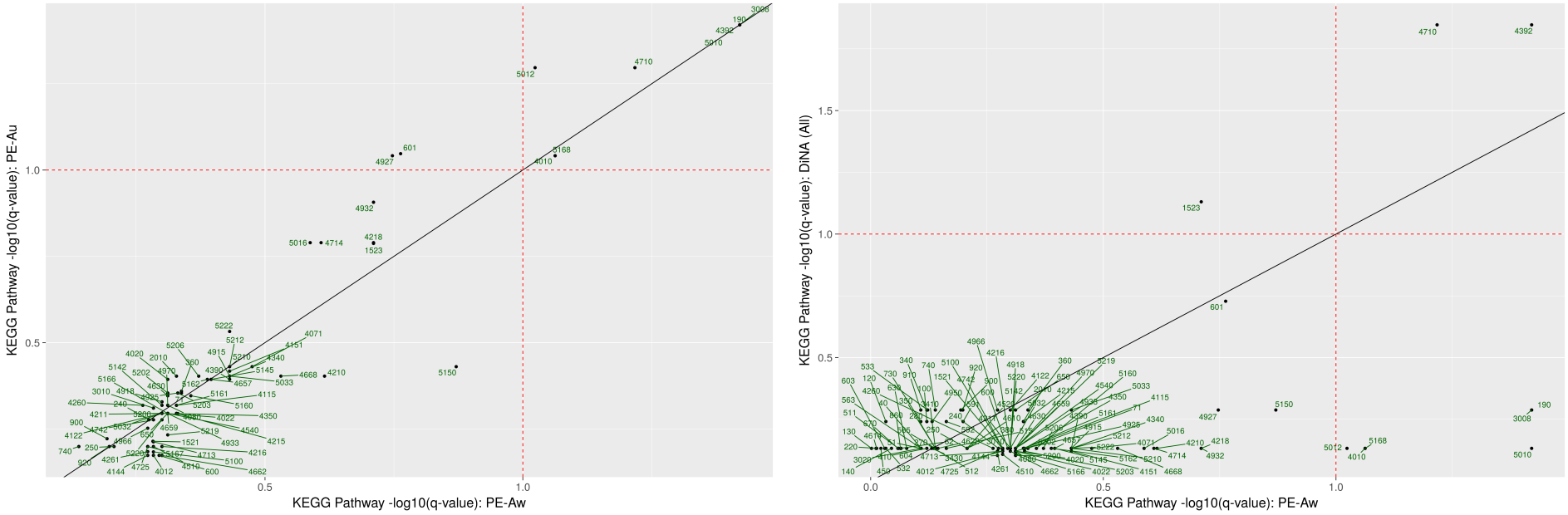
Figure 3.8: **Significant, $q < 0.1$, Pathways by Method for the AEN.** This figure shows the q-values generated by the entropy based methodologies for the AEN for each tested KEGG pathway: PE-A, weighted and unweighted, in addition to DiNA. We can see that the PE-A approaches produce a wider range of pathway enrichments than DiNA, with the addition of weighting information in PE-Aw resulting in the loss of enrichment of two pathways with lower co-expression between pathway members. The 'topology' methods are not shown as they did not have any enrichments that passed the significance threshold. The x axis consists of KEGG pathway ids. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

Investigation into Differential Pathway Enrichment

In order to more thoroughly compare the entropy based methods, it is important to consider the pathways that are shown as enriched by some and not others. In this section I will highlight key differences in results between methods on a pair-wise basis by paying particular attention to pathways that are scored with high significance in one method and low by another. In this section I will primarily be limiting my focus to the top quartile, by significance, of the results for each tool as the pathways contained here will be those deemed most pertinent by each methodology. As the PCN produced a small quantity of q-significant results for DiNA only, I shall here focus on the results for the AEN.

Between the PE-A approach weighting variants, Figure 3.9A shows that many of the more significant pathways in the unweighted method are perhaps overestimated in significance due to their comparative performance when edge weights are used, notably 4927 ('Cortisol synthesis and secretion') and 601 ('Glycosphingolipid biosynthesis - lacto and neolacto series') which rise above a q-value of 0.1 when weights are taken into account. The majority of exceptions to this trend fall above the significance threshold and even then are comparatively few in number, with those of note being pathways 4010 ('MAPK signaling pathway') and 5168 ('Herpes simplex virus 1 infection') which are underestimated by PE-Au. Pathway 4010 known to activate in response to oxidative stress.

Whilst the DiNA method did produce some q-significant results, it produced less p-significant results than the Pathway Entropy methods and as a result the small variety in q-values for its enrichments can be seen in Figure 3.9B. We can note that there is only one differential call by DiNA over PE-Aw: 1523 ('Antifolate resistance') whose relevance to the experimental context is not clear. Those majority of those shown to be differentially significant by PE-Aw can be seen to be both more relevant to the data (190 'Oxidative Phosphorylation', 4010 'MAPK signaling pathway') and those who contain nested portions of these pathways (5010 'Alzheimer's Disease, 5012 'Parkinson's Disease').



(a) PE-A Weighted Versus Unweighted

(b) PE-Aw Versus DiNA

Figure 3.9: **Correlation of q-value for PE-Aw vs PE-Au & PE-Aw vs DiNA.** (A) Pathway enrichment between PE-Aw and PE-Au can be seen to be closely correlated with the addition of network weighting information demonstrating to be a discriminator for the enrichment of 2 pathways. (B) Comparison between PE-Aw versus the DiNA method shows much greater difference in enrichment with DiNA exhibiting a bias toward lower -log10(q) values. Q-values were calculated through the application of Benjamini-Hochberg MTC. The closer to the $x = y$ line the pathway is the more similarly significant it is for both methods. Each point represents a pathway and is labelled with the appropriate KEGG database identifier. The red dashed line indicates the log scaled q-significance threshold: 0.1.

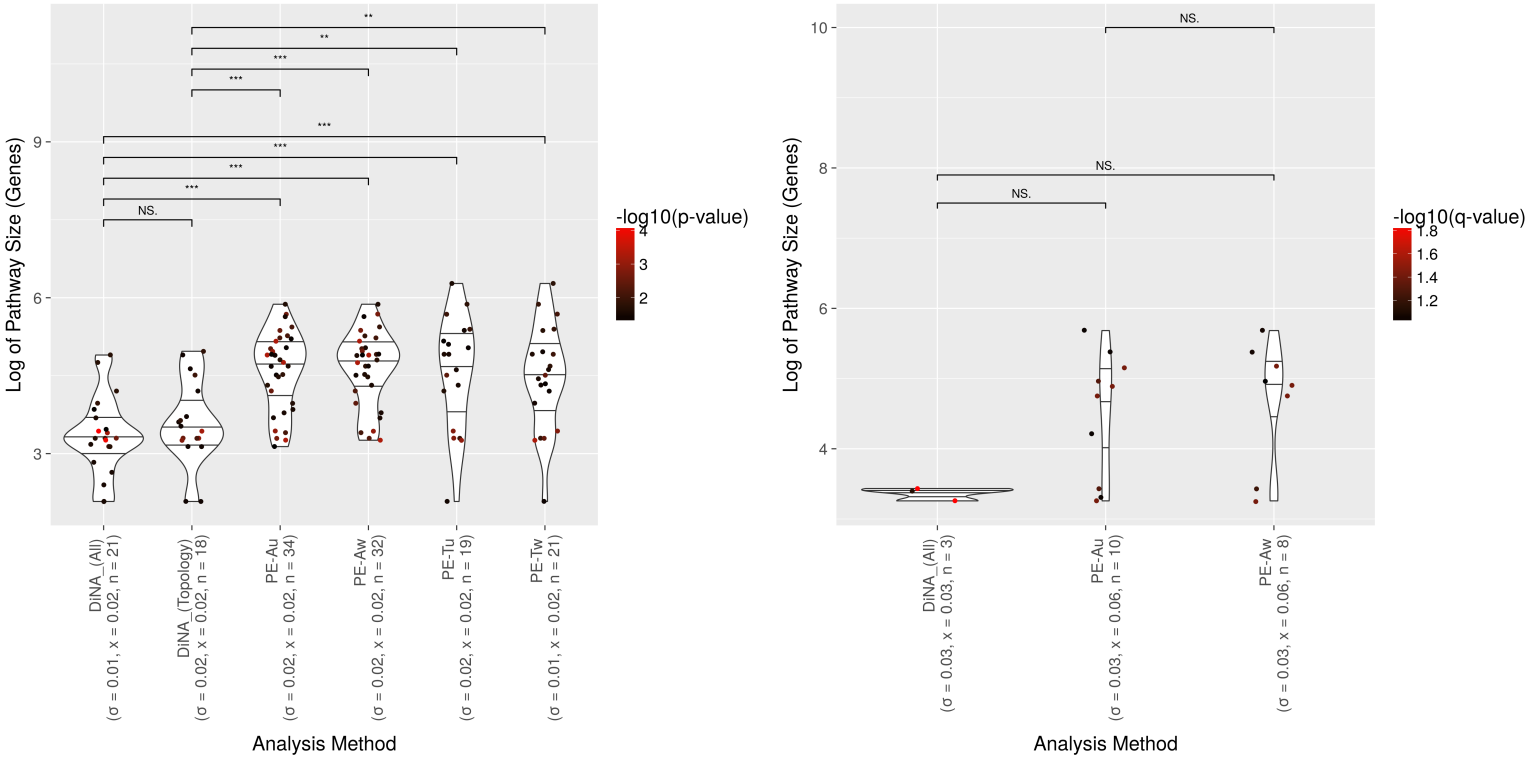
Pathway Size Bias

The pathways in the KEGG database range in size considerably from hundreds of gene members to single digits. We conducted an investigation to determine whether or not the size of a pathway has an impact on the enrichment generated. The results shown here are for the AEN as a result of the PCN not producing q-significant enrichments.

When we visualise only significant ($p < 0.05$) pathways, Figure 3.10A, the DiNA implementations demonstrates a significant bias, as measured by a Wilcoxon Signed-Rank test, towards the significant enrichment of smaller pathways with its enriched pathways demonstrating a smaller overall range that does not encompass larger pathways. Comparatively the PE-T and PE-A methods significantly enrich pathways across a full range of pathway sizes.

We can see in Figure 3.10B for q-values however that this phenomena, whilst suggested, is not significant. This is likely at least in part a result of low statistical power due to the very low number of q-significant pathway enrichments produced by DiNA.

That the Pathway Entropy results do not display the same bias toward the smaller pathways as the DiNA methods, at the level of p-significance, demonstrates the effect of the normalisation for pathway size I have added to the Pathway Entropy methodology.



(a) Significant, $p < 0.05$, Pathways

(b) Significant, $q < 0.1$, Pathways

Figure 3.10: **Method-wise Violin Plots of Pathway Significance Versus Log of Pathway Size, for All Entropy Methods.** At the level of p-significance (A), the DiNA approaches can be seen to demonstrate a significant bias toward smaller pathways. However at the level of q-significance (B) this difference is not preserved. The y axis displays the log of pathway size with the x axis consisting of the tools used and associated vital statistics regarding the significant p and q-values generated: standard deviation in p or q-value (σ), mean p or q-value (x) and number of significant enriched pathways (n). Difference in pathway size distributions was tested using the Wilcoxon signed rank test. The shape of the violin plots is the kernel density of the pathway sizes, with the individual points representing each pathway coloured by p and q-value.

3.3.5 *Evaluation of Network Construction Parameters*

A key motivation for the creation of the Pathway Entropy methods was to explore their use as a tool for evaluating network construction parameters. As the methods have been demonstrated to be resilient to noise by the investigation in Section 3.3.3, the effect of parameter choice should not mask the detection of key pathways, but we might expect choice to affect the enrichment overall. In this section I describe the initial investigation into the evaluative use of the Pathway Entropy methods for two key clustering variables: minimum cluster size and 'deepSplit'.

3.3.5.1 *Minimum Cluster Size*

The minimum cluster size parameter is used by the hierarchical clustering method within the WGCNA tool to determine the minimum size, in genes, for a co-expressed sub-network. As the entropy method relies on assessing how ordered a pathway's representation is in each cluster, this parameter would be expected to have a significant impact on the Pathway Entropy methods' enrichment results. The WGCNA usage instructions recommend a minimum cluster size (MCS) of 30 to be used for analysis and this recommendation has been followed for the analysis thus far in this chapter. As a motivating idea for the creation of the Pathway Entropy methodology was that we might expect genes relating to a certain function, a pathway say, to be similarly co-expressed, I decided to use the smallest value for MCS that WGCNA would accept in order to remove limitation on the grouping of co-expressed genes as many pathways are less than 30 genes in size. Using the experimental dataset described in Section 1.4.1, an empirical test determined that a MCS value below 10 would produce more clusters than WGCNA could handle resulting in error. A MCS of 10 was thus used for comparative evaluation in this section. The significant results for the MCS 10 investigation were calculated using new null distributions, generated using the same procedure detailed in Section 3.2.2.4, however incorporating the decrease in MCS into the generation of the permutation bootstrap networks. Both AEN and PCN networks were explored with MCS 10 and enrichments from both networks are evaluated in this section.

The results for the MCS 10 enrichments for the entropy methods can be seen in Figures 3.11 & 3.12. We can see that for both the AEN and PCN that the largest number of enriched pathways is produced by the PE-A approach, with only the DiNA approaches otherwise producing q-significant enrichments for the AEN. This latter result is interesting as DiNA produced 2 q-significant enrichments for the MCS-30 PCN in Section 3.3.4.1. For the AEN, as one might expect, the change in cluster sensitivity can be seen to reproduce the enrichments from MCS-30 with several additional

pathways now significant. Comparatively the PCN now produces a full set of enrichments for the PE-A approach, perhaps indicating that the increase in cluster sensitivity has compensated for the decrease in network complexity resulting from the removal of the non-coding genes. For both networks, the unweighted PE-A method produces marginally more enrichments than the weighted PE-A, as was seen for the MCS-30 AEN, with a small amount of differentially enriched pathways between the methods.

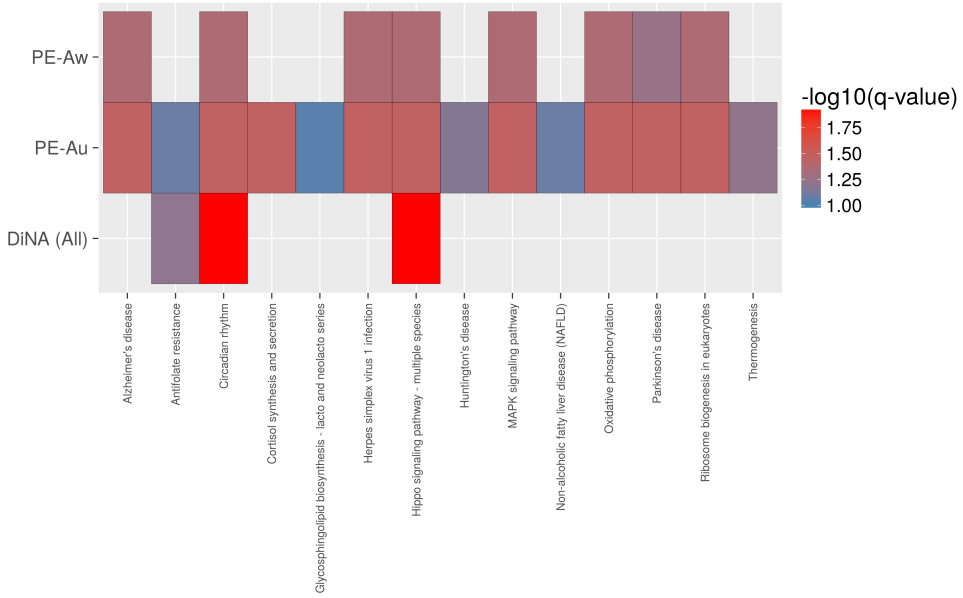


Figure 3.11: **Significant, $q < 0.1$, Pathways by Method for All Entropy Approaches for the AEN, MCS 10.** This figure shows only the significant q-values, $q < 0.1$, generated by the six different entropy based methodologies on the AEN for each tested KEGG pathway for a MCS of 10. As for the encrichment results for a MCS of 30 on the AEN, we see only significant results produced by the PE-A and DiNA approaches, with a wider range produced by the former. The x axis consists of KEGG pathway ids. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

Given the changes in enrichment observed through the change in MCS, I thought it pertinent to inspect how the most significant pathways in the MCS 30 results changed when MCS was set to 10. As AEN produced q-significant enrichments for both MCS thresholds the following figures compare the top quartile of q-significant results for each method for MCS 30 plotted against the same methods' results for MCS 10. This is followed by a comparison of AEN and PCN enrichments for the PE-A approach.

We can see in Figure 3.13A that the decrease in MCS threshold has primarily increased the sensitivity of PE-Aw, but not resulted in a change in the pathways significantly enriched. Comparatively in Figure 3.13B we do see a difference in significant enrichment between MCS levels for the PE-Au method. Of the 4 pathways newly significant, 3 contain a representation of the oxidative phosphorylation pathway (4714

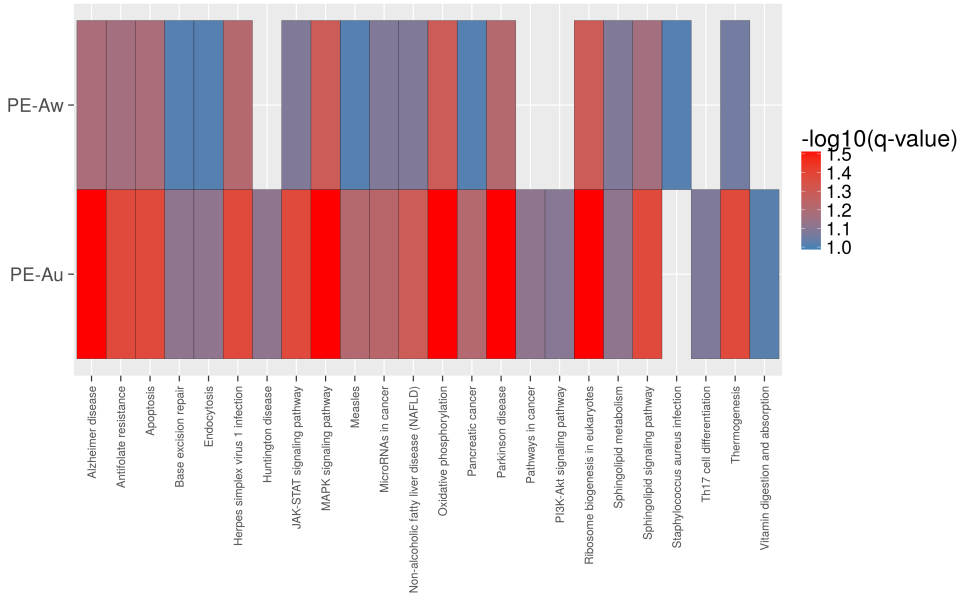


Figure 3.12: **Significant, $q < 0.1$, Pathways by Method for All Entropy Approaches for the PCN, MCS 10.** This figure shows only the significant q-values, $q < 0.1$, generated by the six different entropy based methodologies on the PCN for each tested KEGG pathway for a MCS of 10. Interestingly, with the increase in network complexity resulting from the decrease in MCS, the PE-A approach can be seen to significantly enrich a greater number of pathways. Also of note is that the DiNA method produces no significant enrichments at all. The x axis consists of KEGG pathway ids. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

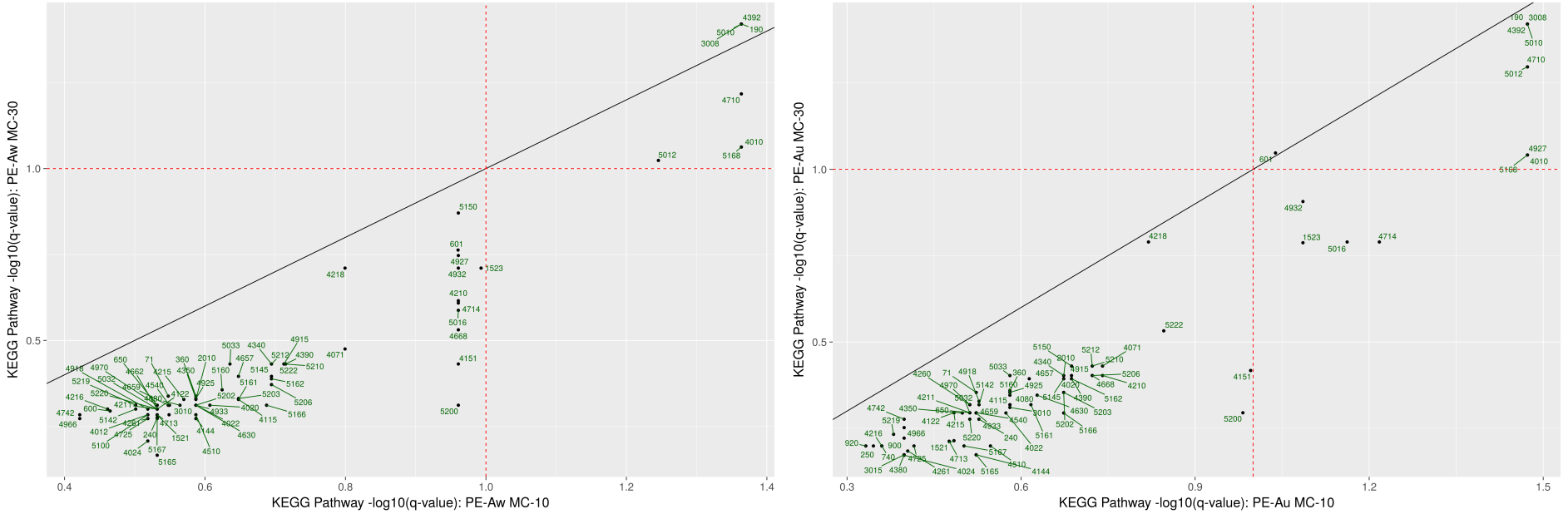
'Thermogenesis', 4932 'Non-alcoholic fatty liver disease (NAFLD)', 5016 'Huntingdon Disease') and the last is a drug resistence pathway whose direct relevance is unclear (1523 'Antifolate resistance'). Thus the decrease in MCS for the PE-A methods in the AEN, whilst increasing enrichment sensitivity, has not produced additional relevant information.

The enrichments produced by the DiNA approach for AEN are entirely consistent between thresholds. For the PCN, we see that a decrease in MCS results in the two pathways identified for MCS 30 no longer significantly enriched. This suggests that the members of these pathways were clustered together in MCS 30 but then split as a result of increased clustering for MCS 10.

Similarly to the AEN network, when we compare the enrichments by the PE-Aw and PE-Au methods for the PCN in Figure 3.14A we see a high level on consensus with only 1 pathway differentially enriched by PE-Aw (5150 Staphylococcus aureus infection). Of the 5 that are differentially enriched by PE-Au, only one has clear relevance to the experimental context (4151 'PI3K-Akt signaling pathway') with one relevent to immune system response but not the cell-types (4659 'Th17 cell differenti-

ation'). Here we can see that the addition of the weighted information provides clear benefit as a descriminator.

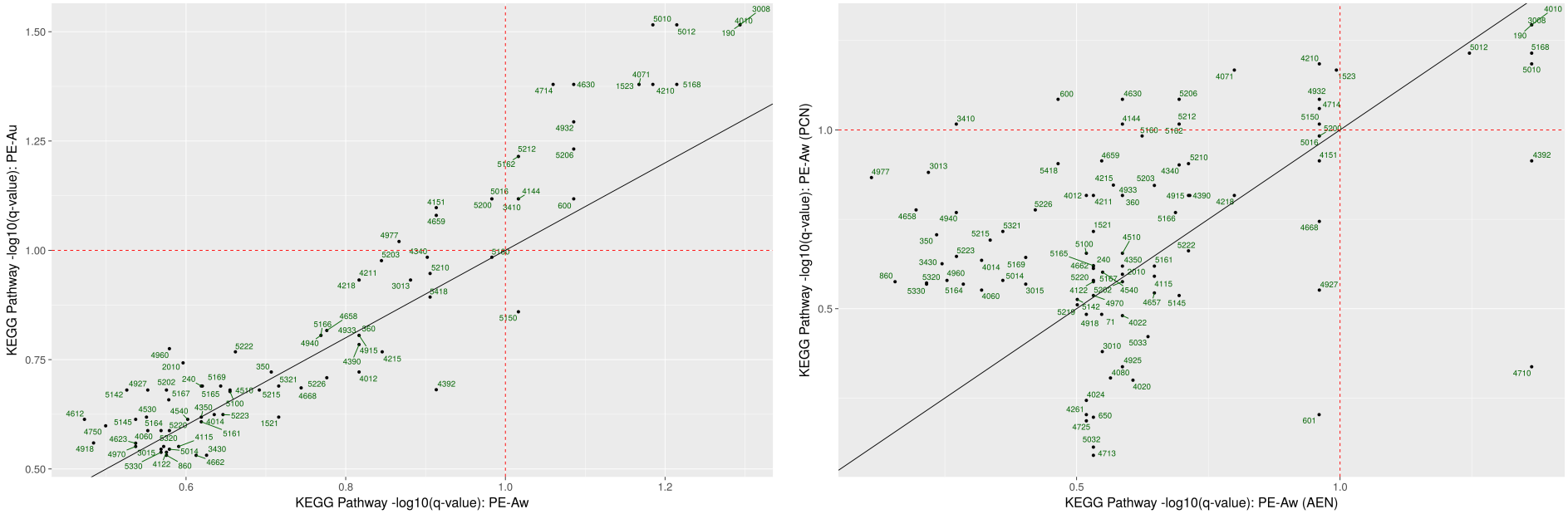
As the PE-Aw method appears, thus far, to produce the most consistent enrichments it is for this method I will compare the enrichment between AEN and PCN networks for MCS 10. If we compare the the performance of this method between networks, we can see in Figure 3.14B that the only pathways differentially enriched in the AEN are the circadian rhythm pathway and the multi species hippo signaling pathways. In addition for the PCN, PE-Aw identifies additional pathways relevent to oxidative stress: one pathway relating to DNA repair (3410 'Base excision repair'), one signaling pathway relating to cell survival (4630 'JAK-STAT signaling pathway'), one to cell death (4210 'Apoptosis') and two sphingolipid pathways (4071 'Sphingolipid signaling pathway' & 600 'Sphingolipid metabolism'). For the latter, the Sphingolipid signaling pathway has been observed to modulate gene expression relating to cell proliferation, differentiation and apoptosis in response to inflammation in glial cells (Colombaioni and Garcia-Gil, 2004). Interestingly the 4144 'Endocytosis' pathway is also differentially enriched by PE-Aw in the PCN, a recent paper has shown evidence for the indirect mediatiation of clathrin-mediated endocyctosis by sphingolipids in astrocytes subject to oxidative stress (Volpert et al., 2017). Lastly there are 5 seemingly less relevent pathways differentially enriched, one infection pathway (5150 'Staphylococcus aureus infection') for which primarily receptor genes are enriched, two pathways containing the oxidative phosphorylation pathway (4714 'Thermogenesis', 4932 'Non-alcoholic fatty liver disease (NAFLD) and two disease pathways containing sub-representations of other enriched pathways (5162 'Measles', 5212 'Pancreatic Cancer').



(a) PE-Aw: MCS10 vs. MCS30

(b) PE-Au: MCS10 vs. MCS30

Figure 3.13: **Correlation of -log10(q-values) for PE-A with Minimum Cluster Size of 10 Versus 30 (AEN).** When comparing the performance of PE-Aw (A) and PE-Au (B) for MCS 10 and 30 we can see a broad increase in sensitivity, however only an increase in enrichment for PE-Au. The newly significant pathways in (B) are also of questionable relevence to the experimental stimulus. Each point represents a pathway and is labelled with the appropriate KEGG database identifier. The red dashed line indicates the log scaled q-significance threshold: 0.1.



(a) PE-Aw (PCN) vs. PE-Au (PCN)

(b) PE-Aw: AEN vs. PCN

Figure 3.14: **Correlation of -log10(q-values) for PE-Aw and PE-Au with Minimum Cluster Size of 10 (PCN & AEN).** (A) PE-Au can be seen to calculate a lower q-value for all those also enriched by PE-Aw, with the use of weighting information in PE-Aw seen to be a descriminator for 4 pathways and indeed producing 1 differential enrichment over PE-Au. (B) Application of PE-Aw to the PCN can be seen to generate a greater number of enrichments whilst preserving the majority identified in the AEN. Each point represents a pathway and is labelled with the appropriate KEGG database identifier. The red dashed line indicates the log scaled q-significance threshold: 0.1.

3.3.5.2 *deepSplit*

The 'deepSplit' variable is a sensitivity parameter for WGCNA's hierarchical clustering method described in more detail in Section 3.2.3.4. The parameter can take integer values between 0 and 4 inclusive with the default value being 2; the values 1 and 0 present a stepped decrease in sensitivity and 3 and 4 a stepped increase. I investigated whether an increase in sensitivity, $\mathtt{deepSplit} = 4$, would result in the more coherent clustering of pathway members.

Having applied this alteration however, the enrichment results for both entropy and hypergeometric results were identical indicating that this sensitivity parameter does not affect the core membership of clustered sub-networks. As a result it can be inferred that the action of this sensitivity parameter must be for the assignment of genes weakly associated with clustered sub-networks.

3.3.6 *Comparative Performance Against Hypergeometric Methods*

In Section 3.3.4.1 I comparatively evaluated my Pathway Entropy methodology both against its predecessor, the only other entropy based approach, and also against variation in the Pathway Entropy methodology itself. In this section I will progress to evaluating my approach against long established methods for pathway enrichment that use a hypergeometric approach: KEGGprofile and clusterProfiler.

3.3.6.1 *Pathway Enrichment*

The hypergeometric results were multiple test corrected using the same method as the entropy approaches: Benjamini-Hochberg MTC. In this section I will evaluate the effect of this process on the hypergeometric results and especially in comparison with the entropy based methods.

The use of only expression information from protein coding genes is essential for this comparison as the methodology of the hypergeometric methods, which test for pathway genes in a group of given genes (a network cluster in our analysis), will be negatively impacted by the presence of genes that are not present in the pathway database: in this case the majority of non-coding genes.

As the removal of the non-coding genome from the network has negatively impacted the sensitivity of the PE methods, a comparison of entropy and hypergeometric methods can be more adequately conducted when the sensitivity of the network construction is increased to compensate for this loss of information. Therefore this comparison analysis will use the MCS 10 thresholded PCN from the previous section

for which the PE-A method has demonstrated performance comparable in sensitivity to its use on the unrestricted AEN.

The enrichment results for all tested entropy methods and hypergeometric methods can be seen in Figure 3.15. We can immediately see that KEGGprofile is perhaps not appropriate for application to this use case as it has determined the majority of the KEGG pathway database to be significantly enriched. clusterProfiler on the other hand enriches a more reasonable quantity of pathways and for which there is a noticeable overlap with the PE-A approach. As PE-A is the only entropy method to produce q-significant enrichments I will focus my comparison henceforth primarily between PE-A and clusterProfiler.

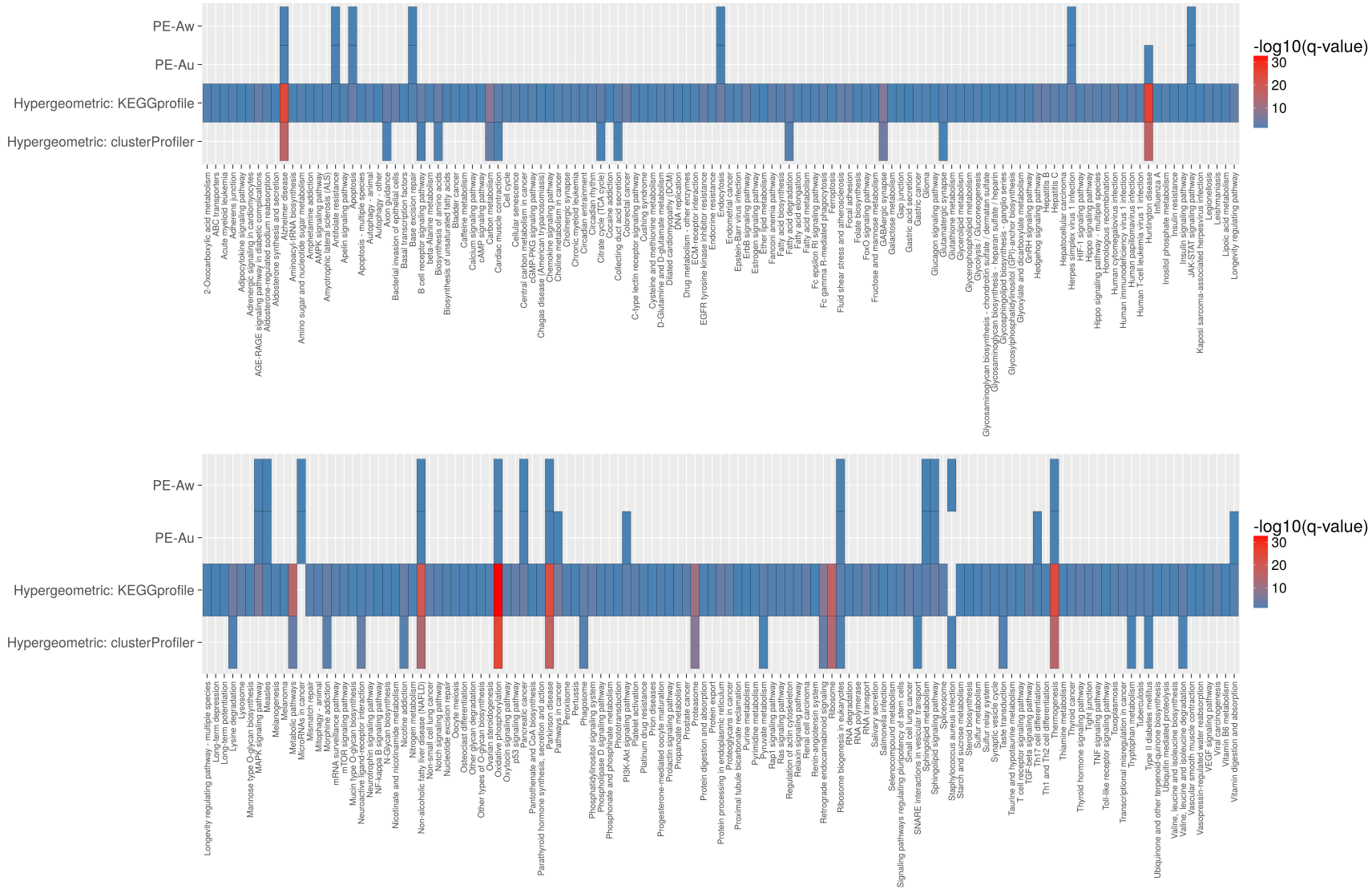


Figure 3.15: **Significant, $q < 0.1$, Pathways by Method for All Entropy and Hypergeometric Approaches.** This figure shows only the significant q-values, $q < 0.1$, generated by the two hypergeometric methods, KEGGprofile and clusterProfiler, in addition to the six different entropy based methodologies for each tested KEGG pathway on the PCN with an MCS of 10. Whilst there is some overlap between PE-A and clusterProfiler (7/23 for PE-Au and 7/19 for PE-Aw), this is less than the difference. KEGGprofile can be seen to determine the majority of the pathways tested as signicant. The x axis consists of KEGG pathway ids. The q-values here have been visualised without upper bound. Missing values for the methods using the 'topology' approach are a result of these pathways lacking defined functional topology in the KEGG database's KGML schema.

3.3.6.2 *Comparative Evaluation*

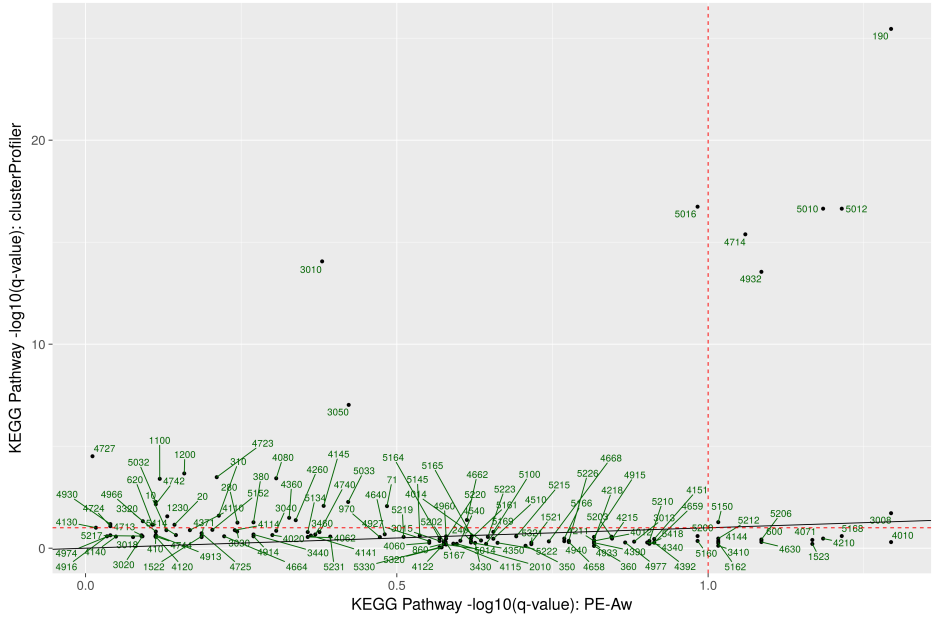
Comparative plots for the top quartile of enrichments, by q-value, can be seen for PE-Aw and clusterProfiler in Figure 3.16A and with PE-Au and the latter in Figure 3.16B.

We can see that, despite the different methodologies, there is a core consensus between the weighted and unweighted PE-A methods and clusterProfiler. All three methods enrich arguably the most relevant pathway: 190 'Oxidative phosphorylation', several pathways for which this pathway is a sub-component (4714 'Thermogenesis', 4932 'Non-alcoholic fatty liver disease (NAFLD)', 5010 'Alzheimer disease' and 5012 'Parkinson disease'), in addition to 3008 'Ribosome biogenesis in eukaryotes'.

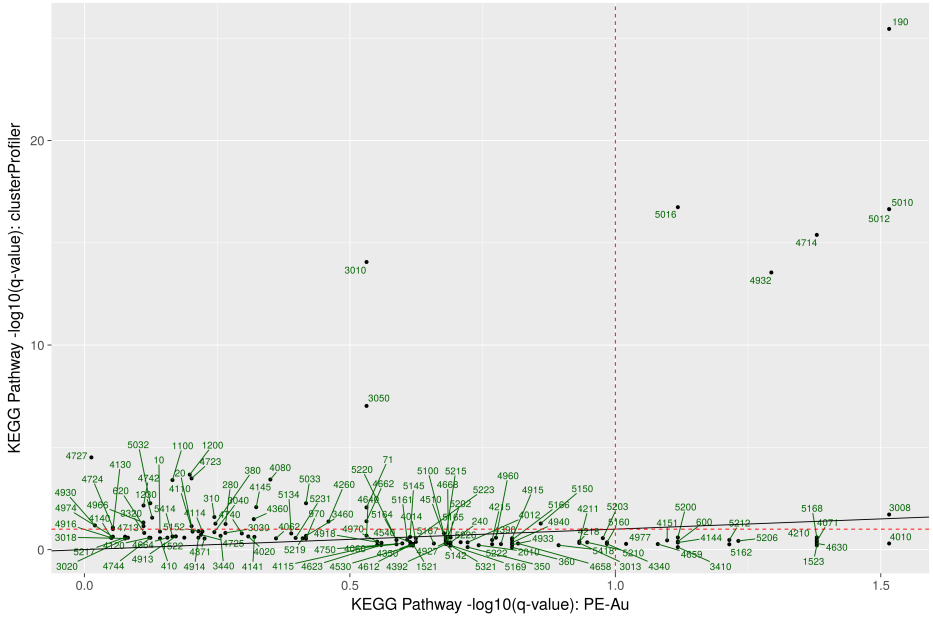
Additionally we can see that PE-Aw and clusterProfiler enrich 5150 'Staphylococcus aureus infection' and that PE-Au and the latter both enrich 5016 'Huntington disease', which has pathway 190 as a sub-component.

Whilst there is a core agreement between these methods, there is a greater disagreement between clusterProfiler and both PE-A approaches. Individually clusterProfiler enriches 25 additional pathways, for which we can see the following specific groups: neural related pathways (4080 'Neuroactive ligand-receptor interaction' ,4360 'Axon guidance', 4723 'Retrograde endocannabinoid signaling', 4724 'Glutamatergic synapse', 4727 'GABAergic synapse' and although not exclusively 4130 'SNARE interactions in vesicular transport'), addiction and disease (4930 'Type II diabetes mellitus', 5032 'Morphine addiction', 5033 'Nicotine addiction'), metabolism (71 'Fatty acid degradation', 1100 'Metabolic pathways', 1200 'Carbon metabolism') for which we have notable sub-groups for both carbohydrate metabolism (20 'Citrate cycle (TCA cycle)', 620 'Pyruvate metabolism') and amino acid metabolism (280 'Valine, leucine and isoleucine degradation', 310 'Lysine degradation', 380 'Tryptophan metabolism', 1230 'Biosynthesis of amino acids').

Whilst the groups mentioned cover the majority of enrichments, there are 7 other pathways whose functions are mostly less directly related to the experimental context, covering functions ranging from sensory to circulation. The exception to this being 4145 'Phagosome' for which the role in inflammation response is relevant.



(a) Pathway Entropy 'all' Weighted vs. clusterProfiler



(b) Pathway Entropy 'all' Unweighted vs. clusterProfiler

Figure 3.16: **Correlation of -log10(q-values) for PE-A and clusterProfiler.** Enrichment attributed by PE-Aw (A) and PE-Au (B) compared to clusterProfiler can be seen to have a small level of of overlap with a large number of mutually differentially enriched pathways. Each point represents a pathway and is labelled with the appropriate KEGG database identifier. The red dashed line indicates the log scaled q-significance threshold: 0.1.

3.3.6.3 *Pathway Size Bias*

As we detected an apparent bias in pathways size during the comparative evaluation of the entropy methods, this analysis has been repeated here to include the hypergeometric approaches.

Figure 3.17 displays a log-scaled violin plot where each data point is the size of a significant pathways, only those methods that produce significant enrichments for the PCN network are displayed. We can see here that the only significant difference between methods, with regards to the size of enriched pathways, is between PE-Aw and KEGGprofile. This is not surprising however when we consider the quantity of pathways KEGGprofile determines to be significant.

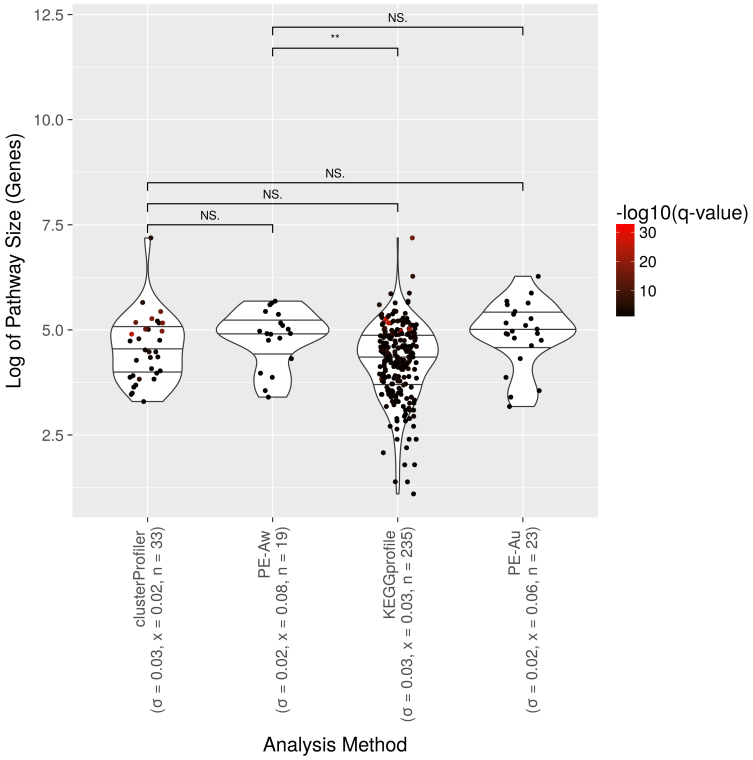


Figure 3.17: **Pathway Significance Versus Log of Pathway Size, for Entropy and Hypergeometric Methods on the PCN.** This figure shows the thresholded q-values for each pathway, ordered by pathway size, generated by the PE-A approaches and the hypergeometric methods for each tested KEGG pathway. The PE-A methods produce enrichments for pathways of a similar size range to clusterProfiler. The only significant difference we can observe is between PE-Aw and KEGGprofile, as the latter determines the majority of tested pathways to be significant, this is not of special interest. The y axis displays the log of pathway size with the x axis consisting of the tools used and associated vital statistics regarding significant q-values generated: standard deviation (σ), mean (x) and number of significant enriched pathways (n). The shape of the violin plots is the kernel density of the pathway sizes, with the individual points representing each pathway coloured by q-value.

3.4 DISCUSSION

Functionality, Performance and Robustness

Methodologically, this project aimed to incorporate structural information from gene co-expression networks within an entropy based framework for the purpose of attributing pathways to clustered gene co-expression networks. The investigations within the Results section of this chapter show that the resultant Pathway Entropy methods have achieved this goal.

With regards to reliability and robustness of results, the p-values are generated using pathway-specific null distributions. Thus a low p-value will necessarily only be generated if the representation and connectivity of a pathway's members is significantly well organised compared to if those genes were randomly ordered within that specific network. Calculating the p-values in this manner reduces the chance of false positives arising from the chance clustering of gene members, especially when the entropy being tested is itself a metric of the degree of organisation of the pathway. Whilst the permutation test used could certainly be improved, and does indeed limit the feature space of p, the values produced have been demonstrated to be stable over changes in the number of permutations and thus present a reliable indication of significance.

Whilst the Pathway Entropy methods certainly generate enrichment results, which have been demonstrated in Section 3.3.3 to be robust to noise, their accuracy with regard to the experimental context of the data is vital for evaluation of the tool's performance. In this regard the methods demonstrated a great sensitivity to the quantity of information in the network. When run on a network constructed from all available gene expression data the PE-A approach returned q-significant enrichments, however when the network's construction was limited to only protein coding gene's expression the same methods only produced enrichments when we lifted the constraint on minimum cluster size: allowing for more clusters to be formed and thus increasing information content. The enrichments produced from both networks were closely overlapping, with a clear majority (77%) of pathways enriched in the PCN being also enriched in the AEN, which we would expect as the Pathway Entropy methods uses only information for the pathway members for its primary calculations, indeed including information from other genes only indirectly through cluster size normalisation. As a result we can assume that the decrease in network size in the PCN negatively impacted the entropy methodology through decreasing the total information in the network, thus highlighting the sensitivity of the approach to information content.

When we consider the nature of the OSm-MA dataset, an astrocyte monoculture subject to oxidative stress (see Section 1.4.1), the pathway enrichment results detailed in Section 3.3.5.1 for both the AEN and PCN show good representation of pathways whose function relates to oxidative stimuli, such as 190 ('Oxidative phosphorylation') and cell death, such as 4210 ('Apoptosis'), for PE-A methods. Indeed signaling pathways associated with oxidative stress or response to stress and inflammation, such as Jak-STAT (4630) and MAPK (4010), are represented well.

Whilst there is a reasonable consensus on the most significant pathway between PE-A weighted and unweighted methods for both networks, there is also observable variation between the weighted and unweighted PE-A, as we might expect upon incorporating more information from the network. Interestingly whilst the PE-T methods produces some of the same enrichments at the p-significance level, however none of these survive multiple testing for either network for both cluster sensitivities, demonstrating perhaps a higher information requirement to satisfy the additional constraints of the approach.

The results of the investigations thus far deal only with one dataset however and so cannot tell us about the performance of the Pathway Entropy methodology in a generalisable manner. In the next chapter, I will be extending the application of this method to additional datasets and so I will leave any conclusive evaluation of the method's performance until then.

The Contribution of Edge Weights & Functional Information

The contribution of the edge weights, the correlation values underlying the clustering of the co-expression network, through the PE-Tw and PE-Aw methods did not contribute as significantly as expected to the enrichment results. Indeed with the addition of weighting PE-T failed to produce significant results. Whilst differences in enrichment were present, as shown in Sections 3.3.4.1 & 3.3.5.1, these differences were reasonably small with the weighted approach seeming to prune several pathways whose underlying co-expression is weaker than the clustering of its members would suggest.

One explanation for the comparably small impact gained through the inclusion of this information is that as it has already been used in shaping the data, its additional utility is limited. As the co-expression networks are clustered on the basis of the edge weights, if a pathway is well represented in a cluster, and so detectable by the unweighted method, then this is a necessary result of the distinct co-expression whose correlation thus facilitated the clustering. If this is the case, then the weighted information would only be informative in situations where connectivity between path-

way members is incidental, for example in large clusters or clusters where genes are shared by a number of pathways in the same cluster.

If this is the case however, there is not much reason not to make use of the edge weight information as its impact, however small, is demonstrable: the apoptosis pathway (4210) is a relevant pathway for the oxidative stress stimulus in OSm-MA, and the addition of the weighted information pushes its p-value up from 0.03 to 0.01 for both PE-Aw and PE-Tw methods for the AEN. Indeed, the difference between the weighted and unweighted PE-A method also bore interesting differential results; for example in the PCN for MCS 10 PE-Au significantly enriches several pathways whose relevance to the data is not directly apparent (4977 'Vitamin digestion and absorption', 5016 'Huntington disease' and 5200 'Pathways in cancer') which are not enriched by the weighted method suggesting that whilst the pathway members are co-expressed, the actual strength of this co-expression is weak and they are not enriched as a result.

The use of functional information, in this case genes involved in a pathway's functional topology of defined links between genes in the KEGG schema, is the core difference between the PE-A and PE-T approaches. As the PE-T approach will only make use of network edges between genes represented in the KEGG schema for each pathway, this represents a substantial constraint on the data that can be used. The effect of this constraint when applied to experimental data, in Sections 3.3.4.1 & 3.3.5.1, has proved to be too great a threshold for information. Whilst the PE-T approach succeeds in demonstrating its ability to detect strong pathway signals, in Section 3.3.3, of the p-significant pathways it produces when applied to experimental data, none pass multiple test correction. As a result it appears that the information requirement for this approach is too high and not appropriate for network-based pathway enrichment as a general use-case.

Comparison to Existing Methods

Established standards for pathway enrichment have been around since before the rise of network based biological analytics and largely follow a hypergeometric based methodology. Since the advent of network biology, and now with the common usage of networks derived from gene co-expression as a tool for scientific enquiry, it still remains common practice to apply such methods either to the data as a whole or a sub-network of interest.

The creation of the DiNA approach by (Gambardella et al., 2013) demonstrated a method by which the same ends could be achieved but instead by rooting the analysis within the context of the data structure being used, thus making for a more robust

approach through the integration of this valuable information. It has been established long ago that structure in biological networks represents biological function (Pržulj et al., 2004). However it has been half a decade since the publication of DiNA and the tool has not caught on and is not actively maintained.

Thus in order to evaluate my improvement upon DiNA's methodology, it is important that we compare not just to DiNA (see Section 3.3.4.1) but also to established representatives of the hypergeometric methods still in common usage (see Section 3.3.6).

Both DiNA and its 'topology' adaptation produce very few q-significant results compared with the Pathway Entropy PE-A approach, for either of the two networks the methods were applied and interestingly for the PCN for MCS 10 we see no significant results. It is curious here that an increase in information content would result in its enrichments for MCS 30 becoming no longer significant, in direct contrast to PE-A for which we see a core robustness in the AEN results and an increase in general sensitivity for all instances applied. The DiNA topology adaptation faired similarly to the PE-T approach, producing no significant enrichments for either network or cluster sensitivity.

The pathways enriched by the DiNA approaches are also mixed in relevancy: for the AEN with MCS 10 we have a signaling pathway relating to cell survival and apoptosis (4392 'Hippo signaling pathway - multiple species') a key pathway for homeostasis (4710 'Circadian rhythm') and a pathway for drug resistance (1523 'Antifolate resistance'). Whilst some of these pathways show a clear relation to the experimental context, they are also not as informative and fewer in number than those enriched by the PE-A approach. It is also worthy of note that we do not see pathway 190 ('Oxidative phosphorylation'), of key relevance to exposure to oxidative stress, in the results of either DiNA approach.

With regard to sensitivity and robustness to noise, investigated in Section 3.3.3, the DiNA methods sensitivity is not significantly different from the pathway entropy apporoches when noise is applied. For higher levels of noise the DiNA approaches can be seen to lose the perturbed pathway signal more quickly. The DiNA enrichment results indicate a bias towards pathways with fewer gene members, this being significant at the p-value level, however not enough enrichments pass multiple testing for this to be confirmed at the q-significance level. Any such bias is alleviated in the Pathway Entropy methods however through the additional normalisation steps.

When comparing the Pathway Entropy methods to the hypergeometric tools KEGGprofile and clusterProfiler in Section 3.3.6, as these methods are not designed to deal with network data I decided to independently apply the hypergeometric tools to each sub-network clustered by WGCNA and take the best result across all the sub-

networks for each pathway. Due to the methodology of the hypergeometric methods a comparison was conducted only on the PCN in order not to risk inflating the test statistics. From the results it was immediately apparent that one of the hypergeometric tools was not adept to such a situation, with KEGGprofile calling almost all pathways tested as significant. clusterProfiler, perhaps by virtue of controlling for gene presence in the network, performed better.

Of the enrichments produced by clusterProfiler, we see that the top 5 most significant pathways are also produced by the PE-A approach, with an additional 2 pathways in common beyond these. Of these common pathways the most significant, as we might expect, is 190 ('Oxidative phosphorylation') and the majority of the others contain a portion of this pathway within them. The exception to this is 3008 ('Ribosome biogenesis in eukaryotes'), which has shown to be upregulated by oxidative stress in studies on yeast (Shenton et al., 2006), and an infection pathway 5150 ('Staphylococcus aureus infection'), however the latter is in common with PE-Aw only. The relevence of 5150 is not explicitly clear, however as it is the genes relating to the cell membrane that form the largest co-expressed grouping we might theorise that the co-expression of these genes could be a result of membrane exposure to the stress stimulus in the astrocytes.

Whilst clusterProfiler and PE-A demonstrate a degree of consensus, there is much greater disagreement between the methods. The pathways differentially enriched by clusterProfiler highlight several functions known to be played by astrocytes in the presence of neurons: modulation of synaptic function via endocannabinoids (4723 'Retrograde endocannabinoid signaling') (Viader et al., 2015), support for axon growth (4360 'Axon guidance') (Anderson et al., 2016) and interaction with glutamatergic and GABAergic synapses (4724 'Glutamatergic synapse', 4727 'GABAergic synapse'), indeed the citrate cycle (KEGG pathway 20) which forms part of the astrocytic role in these processes is also enriched. The significant enrichment of these pathways for astrocytes in the absence of neurons is unexpected and for some could perhaps be explained by the continued function of these roles regardless of neuronal presence. For example a study which cultured *in vivo* astrocytes *in vitro* with microglia has also shown that astrocytes excrete GABA into the surrounding intercellular medium (Lee et al., 2011). Enrichment of the transport pathway 4130 ('SNARE interactions in vesicular transport') perhaps lends tentative support to this hypothesis. In addition to these neuronal pathways, clusterProfiler also enriches several metabolic pathways in keeping with the known function of astrocytes as providers of metabolic support (Bélanger et al., 2011).

When we look at the pathways differentially enriched by the PE-A approach, we can see that they are much more related to the experimental stimulus and resultant

responses to inflammation (4010 'MAPK signaling pathway', 4071 'Sphingolipid signaling pathway', 4630 'JAK-STAT signaling pathway') and cell damage (3410 'Base excision repair', 4210 'Apoptosis'). The role of astrocytes in inflammation response is indeed well characterised (Colombo and Farina, 2016), and the role of sphingolipids in astrocytic response is strongly enriched by the PE-A approach with both signaling and metabolism (KEGG pathway 600) pathways significantly enriched. Interestingly evidence of the indirect mediatiation of clathrin-mediated endocytosis by sphingolipids in astrocytes subject to oxidative stress (Volpert et al., 2017) is also present through the enrichment of the endocytosis pathway.

Thus we can see that whilst a hypergeometric approach highlights the well known roles of astrocytes in support and signaling, it does not highlight pathways indicative of a cellular response to a inflammatory stimulus or indeed to cellular stress as we would expect, beyond the enrichment of the oxidative phosphorylation pathway that is not unique to this approach. Indeed clusterProfiler's strong enrichment of pathways relating to the synapse is also unexpected for an astrocyte monoculture. As a result we can see that the enrichment results of the PE-A approach are most appropriate to the experimental context of the source data.

Pathway Entropy Methods as Evaluative Tool

A limited investigation was carried out into the use of the Pathway Entropy methods as a means of evaluating methodological and parametric decisions for network construction. A more in depth study was prohibited only by project time-constraints. Given the time limitations, two sensitivity parameters for the WGCNA clustering procedure were chosen for study: Minimum Cluster Size and deepSplit.

The deepSplit investigation, detailed in Section 3.3.5.2, demonstrated no change to the highly enriched pathways, thus implicating the sensitivity parameter in cluster precision, where the attachment, or not, of weakly associated genes would not affect the signal of strongly co-expressed genes.

The MCS investigation, detailed in Section 3.3.5.1 was more fruitful in producing a difference in enrichment. The theory behind using the minimum size for MCS was the idea that if, so far as is possible, we do not threshold the cluster sizes then we may see more small and functionally distinct clusters which would, as a necessary result, increase the signal of the pathway(s) represented within them for the Pathway Entropy method.

The results of this application demonstrate an increased sensitivity for both networks to which it was applied, the AEN and the PCN, in terms of both pathway significance metrics and number of significant pathways enriched. However for the

AEN this change primarily affected the PE-Au method, but did not increase the number of enriched pathways relevent to the experimental context. Importantly for the PCN, the decrease in MCS threshold allowed for a sufficient level of information for the PE-A approach to produce enrichment results. Indeed, those enrichments produced by PE-Aw for the PCN network at MCS 10 having the most relevance to the experimental context.

Lastly, a key point of interest with regards to network construction evaluation is the effect of network complexity on enrichment. We can see that the AEN is sufficiently complex for the PE-A approach to yield enrichment results and that those produced for MCS 30 persist at MCS 10, however when we remove the non-coding genes to make the PCN the decrease in network complexity is prohibitive to the function of PE-A. This is perhaps unsurprising as the entropy based methods necessarily incorporate cluster number and size into enrichment calculations. When network complexity in the PCN is increased through the lowering of the MCS threshold, we can see not only that there is now sufficient complexity for PE-A but also that we now see an increase in the relevance of the pathway enrichments, as can be seen in the inclusion of the sphingolipid inflammation response that was not detected in the AEN.

Overall these results demonstrate that decreasing MCS leads to a net increase in enrichment sensitivity as a result of increased information in the clustered gene co-expression networks and that a certain level of information content is required for the performance of entropy based methods. Comparative analysis of the two parameter conditions here has demonstrated that MCS 10 would be a better option to use in cases where the level of information in the network may be less than desired, however this single demonstration of Pathway Entropy as an evaluative tool is not enough to draw conclusions about its evaluative use more broadly. It is thus a motive, and positive indicator, for future investigation into both network evaluation and determining the ideal amount of network information for the use of the Pathway Entropy PE-A approach.

3.5 FUTURE WORK

Throughout the course of the Pathway Entropy project, several plans and ideas for improvement were noted that I hope to implement in the future. These improvements are either drawn from observations made during the testing and evaluation process or present logical next steps for the development of the methodology.

3.5.1 *Implementation of Alternative Significance Methodology*

The permutation test used by the Pathway Entropy method and detailed in Section 3.2.2.4 was utilised as a means for testing the overall functionality of the Pathway Entropy methodology and so as to allow direct comparison with DiNA. As it was chosen for reasons of expedience, it is likely not the most optimal method for significance calculation and thus the implementation of an alternative method is a logical step to reduce the computational load of the method.

3.5.2 *Application to Network Evaluation*

Whilst the Pathway Entropy was applied to the evaluation of network construction parameters during this project, as detailed in Section 3.3.5, this application was limited to sensitivity parameters of existing components of the standard WGCNA network construction methodology.

I had initially hoped to expand this application to included alternative methodologies for network clustering and to comparatively evaluate the effects of the choice of correlation method on the pathway enrichment. Critical evaluation of the WGCNA methodology has started to be published with an alternative k-means clustering procedure proposed by (Botía et al., 2017), however whilst this research seeks to improve the robustness of WGCNA it does not so thoroughly explore the downstream effects of this component of the network construction process.

This investigation would also serve a dual purpose to determining the ideal level of complexity or information in the networks to which the Pathway Entropy methods are applied.

3.5.3 *Extension of Support to Reactome Pathway Database*

For this project I chose to use the KEGG pathway database as it is a long trusted source of pathway information. Its structure is also simple and allowed for expe-

dient integration into my methodology. The newer Reactome pathway database is increasingly more detailed and provides a more complex structure that, whilst more difficult to integrate, can allow for a more detailed enrichment procedure. It is also unrestricted in its availability.

The hierarchical structure of Reactome pathways could allow for differential enrichment of pathway subsections. This would enable a more intelligent normalisation based upon overall versus modular pathway representation, preventing the strong representation of pathway subsections resulting in the enrichment of irrelevant pathways. More informative enrichment results could also be achieved by detailing for what levels of hierarchy a pathway is well represented in a given dataset; if the metabolic portion of a pathway is highly co-expressed, but the rest of the pathway is not, it would be more useful to return a result that describes the co-expression of this pathway component rather than the pathway as a whole.

Thus the extension of support for the Pathway Entropy method to the Reactome database could prove a worthy endeavour.

3.5.4 *Addition of Weighting for Functional Knowledge Representation*

During this project I have trialled and compared two variants of my Pathway Entropy methodology: PE-A and PE-T. These variants emphasise overall co-expression activity and the representation of well characterised functional knowledge respectively.

Given that the additional constraints in the PE-T methods proved too strong for its viability, it is conceivable that one could integrate the idea behind it into the PE-A approach. This could be investigated by weighting the entropy result for a pathway with respect to the representation of explicit functional topology and edges between genes present within it. The exact means by which such a normalisation would best be implemented would require further investigation, however I would posit that the result could be a beneficial synthesis of the two separate approaches.

3.5.5 *Enrichment Accuracy Verification with NiGO*

Whilst most of the pathway enrichments produced by the Pathway Entropy methods are relevant for the datasets they are applied to, this is not universally the case with some seemingly functionally irrelevant pathways also being enriched. An expedient means by which to evaluate the reason for a pathway's enrichment would be to map the genes in the largest co-expressed cluster of gene members to a gene ontology. As the datasets we have applied Pathway Entropy to are neurological in nature, NiGO

(Geifman et al., 2010) would be a good choice as we would be limiting domain information to only what is applicable.

Upon mapping the genes to the ontology, the pathway's relevance can be evaluated by the key functions reported back; for example if we have a disease pathway, however this is only enriched as a result of shared metabolic function, this analysis would be able to report back 'metabolism' as a key function and attach it to the enrichment results to allow for more informed interpretation of results.

3.5.6 *Streamlining Computational Efficiency*

The Pathway Entropy method presently requires reasonable computational resources and significant time to be applied to a dataset. The majority of this is taken up by the permutation bootstrapping process by which the null distribution for significance testing is generated. This process can be conducted in parallel, however as each iteration takes approximately 4 minutes to generate and write to file, discounting the initial network construction, even with ten or twenty threads this task still takes a matter of hours.

An important future task would thus be to investigate alternative means of significance testing or indeed of permutation testing in order to minimise overall time spent on this task. If such an alternative cannot be found then it will be vital to increase the computational efficiency of the process as it is currently implemented in order to achieve the same ends.

3.6 AVAILABILITY

The source code for the Pathway Entropy methods is available on the digital media submitted with the thesis where its location and organisation is listed in Appendix A. As the methodology is still under development it has not yet been made public or packaged up in a form that is easy to use; however updates will be made as development continues.

4

APPLICATION OF SARGASSO AND PATHWAY ENTROPY METHODS TO NOVEL DATA

4.1 MOTIVATION

Having introduced the two novel methods I have worked on during my PhD, Sargasso in Chapter 2 and Pathway Entropy in Chapter 3, I apply them here to further investigate NCAE between neural cell-types. In this chapter I will detail the application of both methods to two mixed-species RNA-Seq datasets: the first is a two species (rat and mouse) investigation into activity dependent changes to the astrocyte transcriptome induced by neuronal function, as published in (Hasel et al., 2017); the second is a three-species dataset investigating the role of microglia on the transcriptome of neurons and astrocytes under inflammatory stimulus, as published in (Qiu et al., 2018). These different datasets belong to the same biological domain, that of neuroscientific study of neurons and glia, thus meaning that any differences we observe in the performance of the methods is likely due to the methods themselves rather than the data. As the same types of cell are used, we can also continue to investigate NCAE between these cell-types, building upon the findings of the previous chapters.

It is important to demonstrate that these two methods are indeed usable and that their performance is adequate across multiple datasets and datasets of different types. In order to emphasise the latter here, particularly with regard to the Pathway Entropy approaches, I have chosen two datasets whose experimental setup differs from the OS dataset used in the previous chapters: one through the structure of experimental conditions and another through the use of three species specific cell-types. The AD dataset, detailed in full in Section 1.4.2, is comprised of a control and two different stimulus conditions, thus lacking the post-stimulus condition of the OS dataset, and the three-species dataset, detailed in full in Section 1.4.3, is comprised of a control and a stimulus condition both conducted with and without the presence of microglia.

These new datasets should thus better demonstrate the efficacy of the novel methods and, importantly, serve to highlight any anomalies in the results of the previous chapters. This investigation also enables us to ask questions regarding the NCAE between cells in this new data. Exclusively the 3Scc dataset, proposes a new challenge for Sargasso in correctly assigning reads in a situation of greater uncertainty

and with a high enough accuracy to enable downstream study. Whilst separation has been demonstrated effectively for two species data, the addition of a third may pose a challenge due to the potential for decreased separability as a result of conservation between the three species. The datasets also propose a challenge to the Pathway Entropy methods in detecting the changes in the affected biological pathways. Whether Sargasso has a detrimental effect on the separation of greater than two species is a point of particular interest. Indeed knowledge derived from these applications will allow for a more thorough evaluation of the performance of these tools.

4.2 APPROACH

In this section, I will present an overview the analysis methodologies for each of the datasets described in the previous section. As both analyses will be using the tools described in the previous two chapters with little deviation, I will keep this overview brief to prevent repetition of methodology.

4.2.1 *Analysis Methodology*

4.2.1.1 *AD Dataset*

The raw RNA-seq data files were mapped to the Ensembl-89 build of both the mouse and rat genome by the `STAR` read aligner, as described in further detail in Section 2.2. I used Sargasso to assign the reads to their species of origin using the information from these alignments using the 'best' filtering strategy, forming two separate species specific, and therefore cell-type specific, datasets for each of the three experimental conditions. The `featureCounts` tool was then applied to these datasets to generate per-gene read counts.

For the application of the Pathway Entropy methodology, recommended pre-processing was first carried out on the per-gene read counts, as described in Section 3.2.3. WGCNA was then used with default settings to construct the gene co-expression network for each species from the respective gene counts, also as described in Section 3.2.3, using counts from only protein-coding genes. To enable permutation testing, 2000 permutations of these networks were made, where for each permutation we randomise the gene labels of the networks adjacency matrix, in order to generate the null distributions for significance testing. The edge lists and clustered modules for each of the species' respective permutations were then saved to file and the null distributions derived to enable Pathway Entropy's application.

After the permutations were created, the species' networks were regenerated using WGCNA, as there is no need to save the network information of the un-randomised network, and the Pathway Entropy PE-Aw method was then used to analyse the representation of KEGG pathways in the clusters of each species' network as it performed best of the variants tested in the previous chapter. Completed enrichment results for the KEGG pathways were then saved to file separately for each species. The graphs in the Results section were then plotted directly from the species-specific enrichment results.

A minimum cluster size of 30 was used both in order to keep parameters in line with the WGCNA author's recommendations and as the higher sequencing depth

and longer reads should provide greater network complexity than that of the OS dataset. Whilst variation in this factor was experimented with in Chapter 3, further work would need to be carried out to solidify any conclusions about the use of the Pathway Entropy methods with a smaller minimum cluster size.

4.2.1.2 *3Scc Dataset*

The analysis methodology for this dataset differs from that for the AD dataset only marginally. I ran Sargasso using the Ensembl-89 build of the human genome in addition to rat and mouse, assigning the reads also using the 'best' filtering strategy. When separating for three species, Sargasso's methodology changes only marginally: each read is separately aligned to each of the three species' reference genomes after which read assignment must satisfy the same criteria in order to assign a read. As such the risk of non-assignment due to ambiguity will potentially be greater, for three species than two, due to a possible increased region of conserved genome between any two or even all three of the three species.

My use of Sargasso to separate the 3Scc dataset thus resulted in three separate, species specific, datasets for each of the four experimental conditions.

After this point `featureCounts`, WGCNA and the Pathway Entropy methods were applied in the same manner as described for the AD dataset in the previous section.

4.2.2 *Considerations*

As both the datasets to be analysed in this chapter differ in some way from the OS dataset that has been used thus far, it is important to note these differences as they may affect the results and our interpretation of them.

Whilst the number of conditions for the AD dataset is the same as that in the OS dataset, it has an additional replicate for each condition. The 3Scc dataset contains 3 replicates for each condition however has an additional conditions compared to the other two datasets used in this thesis. This extra condition and replicate represent a greater depth of information for these two datasets that has the potential to result in a more defined co-expression networks, which may affect the definition of their clustering. Whilst the additional replicate for the AD dataset should not present a substantial change to the overall enrichment, as the co-expression is after all measured on change between conditions, the extra condition for the 3Scc dataset may result in a potentially more detailed enrichment as a result of this extra data point for co-expression.

For the 3Scc dataset, there will necessarily be a higher region of inseparability between the mouse, rat and human genomes as a result of evolutionary conservation

between the species. As this conservation only needs to take place between two of the three species for a read to be rejected (in cases where it can not be unambiguously mapped to the third), there are now three potential reasons for species similarity based read rejection: conservation between species 1 and 2, 1 and 3 or indeed all three. The increase in read length however, should serve to reduce the probability of these occurrences as there will be less instances of perfect conservation of 75bp over 50bp.

4.3 RESULTS

4.3.1 *AD Dataset Results*

4.3.1.1 *Species Separation*

The Sargasso separation of this dataset has been described previously in Section 2.3.9.1 for Sargasso's 'conservative' filtering strategy. For the network construction, the 'best' filtering strategy was used to minimise read loss, the results for this assignment can be seen in Table 4.1. The protein coding read assignment for this strategy can be seen in Table 4.2. These results are as expected considering the assignment in 2.3.9.1.

ADcc 'best'		Mouse				Rat			
Condition	Replicate	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous
Control (w/ TTX)	1	107566855	47192128 **(43.9%)**	57573943 **(53.5%)**	2800784 **(2.6%)**	143570936	108830187 **(75.8%)**	31939965 **(22.2%)**	2800784 **(2.0%)**
Control (w/ TTX)	2	69523733	19084511 **(27.5%)**	48420130 **(69.6%)**	2019092 **(2.9%)**	114499510	94536711 **(82.5%)**	17943707 **(15.7%)**	2019092 **(1.8%)**
Control (w/ TTX)	3	111421040	43524949 **(39.1%)**	64849842 **(58.2%)**	3046249 **(2.7%)**	159347528	124954497 **(78.4%)**	31346782 **(19.7%)**	3046249 **(1.9%)**
Control (w/ TTX)	4	74915044	28389982 **(37.9%)**	44455180 **(59.3%)**	2069882 **(2.8%)**	108565183	85546222 **(78.8%)**	20949079 **(19.3%)**	2069882 **(1.9%)**
Bicuculine	1	105999393	47333406 **(44.7%)**	55913058 **(52.7%)**	2752929 **(2.6%)**	141633485	106691105 **(75.3%)**	32189451 **(22.7%)**	2752929 **(1.9%)**
Bicuculine	2	85170094	24459822 **(28.7%)**	58204555 **(68.3%)**	2505717 **(2.9%)**	140810255	116114649 **(82.5%)**	22189889 **(15.8%)**	2505717 **(1.8%)**
Bicuculine	3	92157237	37027463 **(40.2%)**	52655805 **(57.1%)**	2473969 **(2.7%)**	131516688	102390746 **(77.9%)**	26651973 **(20.3%)**	2473969 **(1.9%)**
Bicuculine	4	73376722	30888653 **(42.1%)**	40526743 **(55.2%)**	1961326 **(2.7%)**	102069648	78583624 **(77.0%)**	21524698 **(21.1%)**	1961326 **(1.9%)**
Bicuculine+TBOA	1	84228968	37751006 **(44.8%)**	44290988 **(52.6%)**	2186974 **(2.6%)**	111951585	84166775 **(75.2%)**	25597836 **(22.9%)**	2186974 **(2.0%)**
Bicuculine+TBOA	2	81256071	26801856 **(33.0%)**	52082792 **(64.1%)**	2371423 **(2.9%)**	125313125	101066398 **(80.7%)**	21875304 **(17.5%)**	2371423 **(1.9%)**
Bicuculine+TBOA	3	100130376	40237520 **(40.2%)**	57283399 **(57.2%)**	2609457 **(2.6%)**	142757122	111111027 **(77.8%)**	29036638 **(20.3%)**	2609457 **(1.8%)**
Bicuculine+TBOA	4	82202145	33098045 **(40.3%)**	46697906 **(56.8%)**	2406194 **(2.9%)**	115704331	89873931 **(77.7%)**	23424206 **(20.2%)**	2406194 **(2.1%)**

Table 4.1: **Sargasso Read Assignment for ADcc.** This table describes the Sargasso assignment results for the ADcc dataset for the 'best' filtering approach. The results from this assignment were used to assess protein coding reads lost as a result of Sargasso's application.

ADcc 'best'		Reads Mapped to Genome		Reads Assigned		% Lost	
Condition	Replicate	Mouse	Rat	Mouse	Rat	Mouse	Rat
Control (w/ TTX)	1	88491963	96359140	40465818	74669299	54.3	22.5
Control (w/ TTX)	2	57702998	77313875	16428489	66187652	71.5	14.4
Control (w/ TTX)	3	92173370	106730734	37084470	85859812	59.8	19.6
Control (w/ TTX)	4	62619129	72784041	24410639	58670972	61.0	19.4
Bicuculine	1	88228447	94405796	40493173	72322729	54.1	23.4
Bicuculine	2	71183555	93119358	20758082	79230135	70.8	14.9
Bicuculine	3	76836105	87717739	31775637	69760892	58.6	20.5
Bicuculine	4	61356985	68376186	26679184	53668055	56.5	21.5
Bicuculine+TBOA	1	70379569	75548422	32453197	57732170	53.9	23.6
Bicuculine+TBOA	2	68099296	85122902	22992832	70596158	66.2	17.1
Bicuculine+TBOA	3	83428538	96181744	34315779	76587793	58.7	20.4
Bicuculine+TBOA	4	68760481	78590612	28351453	62154534	58.8	20.9

Table 4.2: **Sargasso Assignment of Protein Coding Reads for the ADcc Dataset**. Sargasso's 'best' strategy assigns protein coding reads for the ADcc with slightly less proportionate loss that for all reads. The figures here describe uniquely mapping reads only, with reads lost here including both rejected and ambiguous reads.

4.3.1.2 *Pathway Enrichment*

The pathway enrichments produced by PE-Aw for the mouse astrocyte (ADcc-MA) species separated dataset can be seen in Figure 4.1. When applied to the rat neuron dataset (ADcc-RN) PE-Aw produced a number of p-significant enrichments, however none of these were determined to be q-significant after the application of BH MTC.

For the ADcc-MA results, we see a reasonably large enrichment set at 35 pathways. Within this we can see several clear areas of activity: metabolism , signaling molecules and transport, immune response and cell death. We also see a reasonable number of disease pathways enriched.

Looking closely at these pathways we see several important pathways relating to the non-cell autonomous functions of astrocytes. With regards to the signaling pathways, TNF signaling (KEGG: 4668) by astrocytes is known to modulate synaptic strength (Stevens, 2008) and release of TNF can be stimulated by NOD2 (Jensen et al., 2013). The cAMP signaling pathway (KEGG: 4024) highlights the activity of the astrocytic portion of the astrocyte–neuron lactate shuttle (Hasel et al., 2017), which results in increased lactate export to support neural metabolism.

Outside of signaling, we see the enrichment of the 'Citrate cycle (TCA cycle)' pathway (KEGG: 20) which plays a key metabolic role in astrocytic support at GABAergic and glutaminergic synapses, where excess GABA and glutamate are cleared from the synapse and processed by astrocytes. Enrichment of the phagosome pathway (KEGG: 4145) may indicate astrocyte phagocytosis in response to the high levels of synaptic

activity promoted by the experimental stimulus (Bellesi et al., 2017). Similarly the presence of the lysosome pathway (KEGG: 4142) is likely a result of the provision of ATP to support neurons via astrocytic lysosomes (Zhang et al., 2007).

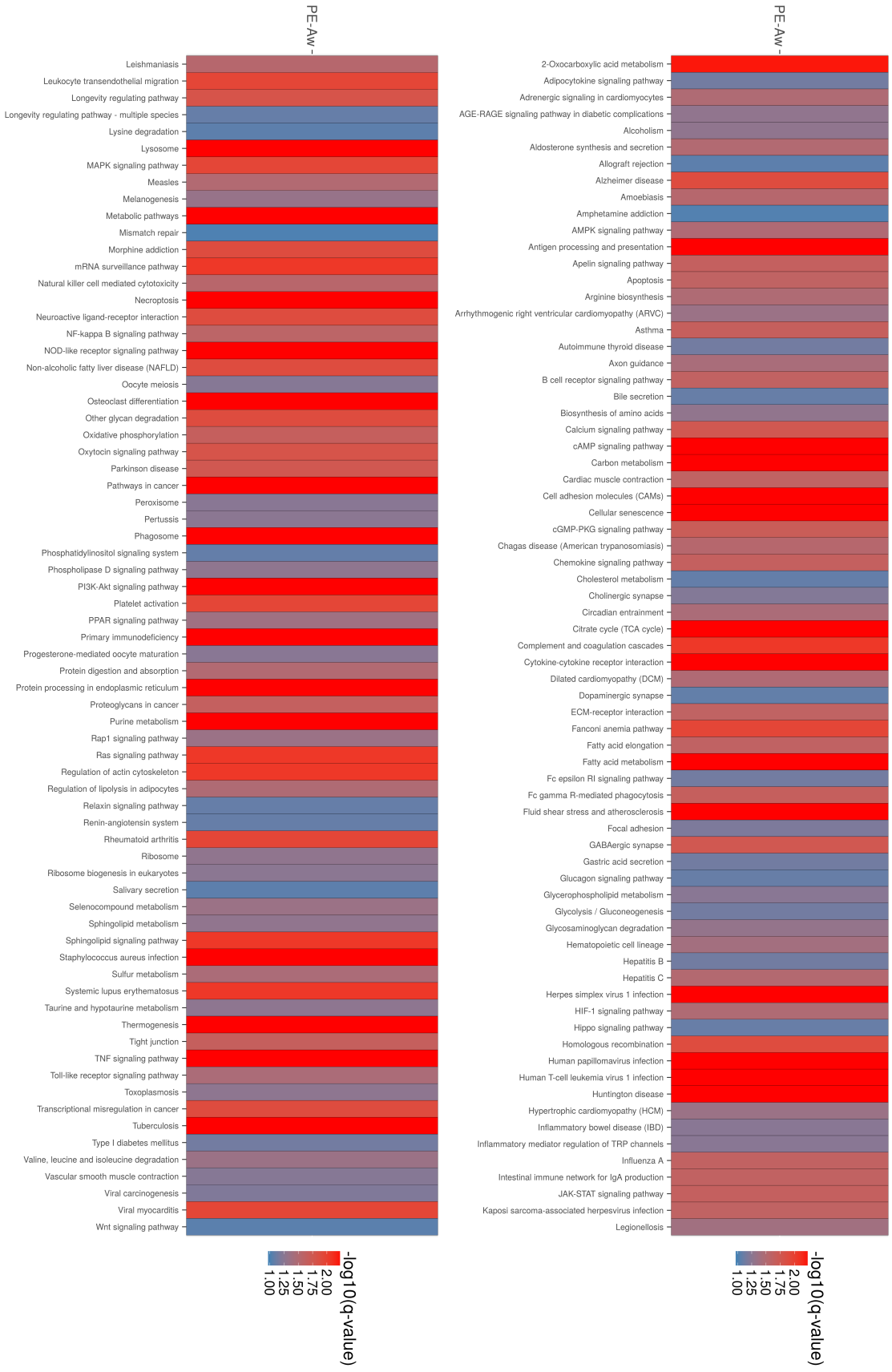


Figure 4.1: **Significantly Enriched,** $q < 0.1$ **Pathways for PE-Aw Applied to ADcc-MA.** The application of PE-Aw to the species separated dataset ADcc-MA uncovers pathways relating to astrocytic function and to the NCAE role of astrocytes. The latter providing additional validation to evidence for the role of cAMP-signaling in the astrocyte-neuron lactate shuttle as reported in (Hasel et al., 2017). Pathways are labelled by name in the KEGG pathway database and q-values have been log scaled.

4.3.2 *3Scc Dataset Results*

4.3.2.1 *Species Separation*

I applied Sargasso to the 3Scc dataset to disambiguate reads from mouse neurons (3Scc-MN), human astrocytes (3Scc-HA) and rat microglia (3Scc-RM). The dataset consisted of 4 conditions, control and LPS both with and without microglia present, with each condition being repeated in triplicate. The conditions without microglia were subject to the same three species assignment so as to produce a consistent separation for the mouse and human data; otherwise regions of ambiguity would change between samples. As such the assignment figures for microglia on these samples give us an indication of the level of mis-assignment for the dataset overall and particularly for the rat. A summary of the assignment can be seen in Table 4.3, with figures for the separation of protein coding reads in Table 4.4.

For the two conditions without microglia, we have a lower read depth of 100 million reads. As me may expect from a three species separation, we have a higher level of ambiguity with roughly 4 million reads that cannot be assigned to each of the three species due to species similarly, a mean of 4.1%, 6.6% and 5.0% of total mapped reads for mouse, human and rat respectively. Despite this ambiguity we can see that the proportion of reads assigned to rat is very small.

The relative proportions of reads assigned to each species reflect difference in population in the experimental co-cultures, with a significantly larger amount of mouse neurons than either glial cell-type and a similar proportion of human astrocytes and rat microglia to one another. Unfortunately replicate matched cell purity information is not available to provide further quantitative validation. However, given that these cultures, particularly the three species conditions, have been sequenced to a high depth, the relative inequality of read totals should not negatively affect the detection and representation of gene expression.

When looking at the mapping totals, we can see that the majority of reads mapped to the human and rat genome are rejected, likely as a result of better mapping to the mouse, with a consistent proportion of ambiguity throughout the 4 conditions and their replicates. We also see a higher number of mis-assigned reads, than observed for the AD dataset in Section 2.3.9, a mean of 1.7% over the six samples, as a percentage of all reads that map to the rat. Interestingly we can see one potentially anomalous result for replicate 2 of the 3Scc LPS condition where we see a drop in proportion of reads assigned to mouse, to 51.8% from a mean of 82.0% across the 11 other samples, and an increase in read assignment to both human, to 30.3% from a mean of 20.8% across the 11 other samples, and rat, to 39.8% from a mean of 12.2% across the other 5 samples containing microglia. Whilst this is not actually the highest proportion

assigned to human, that being in replicate 1 of the two species LPS condition, this is quite an increase for the microglia. We also see somewhat of an increase in read assignment to 3Scc-RM in the 3rd replicate of the 3Scc LPS condition, with also less assigned to 3Scc-MN, however assignment for 3Scc-HA isn't any higher here. As the pattern of fluctuation in read assignment here is not consistent across samples, it is possible that cell population proportions in these particular co-cultures may have been different than in other samples.

With regard to the separation of protein coding reads, we can see that comparative proportion of protein coding reads lost is actually less than for the total reads. This demonstrates that the addition of a third species to the separation has not affected the capability of Sargasso to correctly assign protein coding reads.

Overall we can see that Sargasso has separated the mixed-species dataset into three species and cell-type specific datasets, these we will use for the subsequent analysis in this section.

		Mouse				Human				Rat			
Condition	Replicate	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous	Reads Mapped	Assigned	Rejected	Ambiguous
Control (A & N)	1	97591073	85017094 **(87.1%)**	8530473 **(8.7%)**	4043506 **(4.1%)**	49477442	6331688 **(12.8%)**	39561839 **(80.0%)**	3583915 **(7.2%)**	78389470	1445726 **(1.8%)**	72958682 **(93.1%)**	3985062 **(5.1%)**
Control (A & N)	2	94658843	81075611 **(85.6%)**	9642300 **(10.2%)**	3940932 **(4.2%)**	50552844	9448676 **(18.7%)**	37598744 **(74.4%)**	3505424 **(6.9%)**	75970465	1332899 **(1.8%)**	70762520 **(93.1%)**	3875046 **(5.1%)**
Control (A & N)	3	108975052	90945097 **(83.5%)**	13549396 **(12.4%)**	4480559 **(4.1%)**	63570282	17898165 **(28.2%)**	41684367 **(65.6%)**	3987750 **(6.3%)**	87625716	1189596 **(1.3%)**	82063546 **(93.7%)**	4372574 **(5.0%)**
LPS (A & N)	1	113419381	97812484 **(86.2%)**	10874267 **(9.6%)**	4732630 **(4.2%)**	59075251	9246208 **(15.7%)**	45621742 **(77.2%)**	4207301 **(7.1%)**	90841780	2019604 **(2.2%)**	84159723 **(92.6%)**	4662453 **(5.1%)**
LPS (A & N)	2	94263282	77838609 **(82.6%)**	12605838 **(13.4%)**	3818835 **(4.0%)**	56272186	16427580 **(29.2%)**	36436305 **(64.8%)**	3408301 **(6.0%)**	76201646	1418139 **(1.9%)**	71027877 **(93.2%)**	3755630 **(4.9%)**
LPS (A & N)	3	104437130	86938835 **(83.2%)**	13334508 **(12.8%)**	4163787 **(4.0%)**	61731394	18991809 **(30.8%)**	39041676 **(63.2%)**	3697909 **(6.0%)**	83252139	1021815 **(1.2%)**	78158765 **(93.9%)**	4071559 **(4.9%)**
Control (A,N & M)	1	226318302	185389555 **(81.9%)**	31482982 **(13.9%)**	9445765 **(4.2%)**	111079674	10200435 **(9.2%)**	92719123 **(83.5%)**	8160116 **(7.3%)**	187375518	21019291 **(11.2%)**	156985491 **(83.8%)**	9370736 **(5.0%)**
Control (A,N & M)	2	207860292	168664819 **(81.1%)**	30442809 **(14.7%)**	8752664 **(4.2%)**	110407078	20141679 **(18.2%)**	82596294 **(74.8%)**	7669105 **(7.0%)**	171086190	15010284 **(8.8%)**	147467999 **(86.2%)**	8607907 **(5.0%)**
Control (A,N & M)	3	194335072	154168994 **(79.3%)**	32250879 **(16.6%)**	7915199 **(4.1%)**	111425134	29373338 **(26.4%)**	75107128 **(67.4%)**	6944668 **(6.2%)**	160704677	13960442 **(8.7%)**	138961899 **(86.5%)**	7782336 **(4.8%)**
LPS (A,N & M)	1	183892731	132927930 **(72.3%)**	42254121 **(23.0%)**	8710680 **(4.7%)**	96380014	11506247 **(11.9%)**	77487150 **(80.4%)**	7386617 **(7.7%)**	160918851	35098371 **(21.8%)**	117188541 **(72.8%)**	8631939 **(5.4%)**
LPS (A,N & M)	2	165219931	85555102 **(51.8%)**	72198649 **(43.7%)**	7466180 **(4.5%)**	105369077	31937844 **(30.3%)**	66902117 **(63.5%)**	6529116 **(6.2%)**	159048156	63268106 **(39.8%)**	88342126 **(55.5%)**	7437924 **(4.7%)**
LPS (A,N & M)	3	281854285	221531192 **(78.6%)**	49304921 **(17.5%)**	11018172 **(3.9%)**	164010630	44476618 **(27.1%)**	109813018 **(67.0%)**	9720994 **(5.9%)**	234443563	24270126 **(10.3%)**	199204169 **(85.0%)**	10969268 **(4.7%)**

Table 4.3: **Sargasso Species Assignment Summary for 3Scc Dataset**.
This table summarises the assignment of reads by Sargasso for the 3Scc dataset, where Sargasso can be seen to have successfully partitioned a dataset of three species RNA-Seq reads derived from three cell-types. The number of reads assigned, rejected and unassigned due to ambiguity are listed for each species for each condition. For the conditions, conditions containing only astrocytes and neurons are labelled "(A & N)", whilst conditions containing astrocytes, neurons and microglia are labelled "(A,N & M)".

3Scc Condition	Replicate	Reads Mapped to Genome			Reads Assigned by Sargasso			Percentage Lost		
		Mouse	Rat	Human	Mouse	Rat	Human	Mouse	Rat	Human
Control (A & N)	1	78553366	53410860	38658394	69190875	501751	5145169	11.9	99.1	86.7
Control (A & N)	2	75777206	51830377	40033817	65517146	461844	7829023	13.5	99.1	80.4
Control (A & N)	3	88144697	61262475	51580120	73993952	485921	14925048	16.1	99.2	71.1
LPS (A & N)	1	90390258	61906694	46170679	78837581	793727	7501161	12.8	98.7	83.75
LPS (A & N)	2	74650769	52060710	44554574	62253573	446123	13601311	16.6	99.1	69.5
LPS (A & N)	3	83855897	58205239	49932217	70100920	446812	15625382	16.4	99.2	68.7
Control (A,N & M)	1	182926786	129598190	88208545	150889314	14026825	8609904	17.5	89.2	90.2
Control (A,N & M)	2	165049281	116426823	86651582	135274250	9656910	16611318	18.0	91.7	80.8
Control (A,N & M)	3	157718254	113040285	90657543	125686500	9670437	24560898	20.3	91.5	72.9
LPS (A,N & M)	1	145450080	110139987	74040803	107051248	23571484	9102237	26.4	78.6	87.7
LPS (A,N & M)	2	129757426	111001511	81730935	68195525	44893170	26364969	47.4	59.6	67.7
LPS (A,N & M)	3	227414316	164839900	132022216	179878639	16710493	36739307	20.9	89.9	72.2

Table 4.4: **Sargasso Assignment of Protein Coding Reads for the 3Scc Dataset**.
Similarly to the Sargasso separation for ADcc, the assignment of protein coding reads here falls slightly short of the proportion of reads lost when all reads are used. For the rat microglia to whom no reads belong in the first two conditions, we see an almost total rejection of all mapping reads for these conditions. The figures here describe uniquely mapping reads only, with reads lost here including both rejected and ambiguous reads. For the conditions, conditions containing only astrocytes and neurons are labelled "(A & N)", whilst conditions containing astrocytes, neurons and microglia are labelled "(A,N & M)".

4.3.2.2 *Pathway Enrichment*

Significant enrichments produced through the application of PE-Aw to the species separated 3Scc datasets can be seen in Figure 4.2.

The 'anticipated results' section in our paper (Qiu et al., 2018) details the results of differential expression performed on 3Scc-RM. One of the repressed genes, Igf1, is an anti-inflammatory marker gene involved in several pathways significantly enriched by PE-Aw with a variety of function: 1522 'Endocrine Resistance', 4014 'Ras signaling pathway', 5410 'Hypertrophic cardiomyopathy (HCM)' and 5414 'Dilated cardiomyopathy (DCM)'. Several of these pathways are also relevant to the experimental context with Ras signaling being involved in cell survival and shown to be induced by LPS stimulation in rat microglia (Gong et al., 2019). The endocrine resistance pathway, whilst not explicitly relevant itself, is significantly enriched here due to the strong representation of the MAPK and PI3K-Akt signaling pathways within it; pathways which deal with pro-inflammatory stimulus like LPS. The relevance of the cardiomyopathway pathways however is not explicitly clear, however they do both contain a representation of the renin-angiotensin (Ras) system. Igf1 is an insulin growth factor and whilst not itself involved in the 4911 'Insulin secretion' pathway, it is perhaps notable that this is significantly enriched. The well induced genes mentioned in (Qiu et al., 2018): Cxcl1-3, Ccl3, Il1a and Il1b, are well-known LPS/TLR4-inducible cytokine/chemokine genes, however the pathways these are known to be involved in are not represented in the significant enrichments of the Pathway Entropy approaches. This is surprising considering that several of these pathways are known to be stimulated in microglia by LPS: 4064 'NF-kappa B signaling pathway', 4620 'Toll-like receptor signaling pathway' and 4621 'NOD-like receptor signaling pathway', and have been observed by a previous meta-study on LPS in microglia using WGCNA (Holtman et al., 2015). However manual investigation of the location of these pathway's gene members in the network clustering showed them to be split amongst either 3 or 4 large clusters, resulting in a poor entropy value.

The microglia enrichment does produce other relevant pathways that do not include these genes: a structural pathway affected by LPS (4540 'Gap junction'), pathways relating to anti-inflammation response (531 'Glycosaminoglycan degradation' (Linnartz et al., 2012), 4915 'Estrogen signaling pathway' (Vegeto et al., 2001), 4921 'Oxytocin signaling pathway' (Yuan et al., 2016), 4925 'Aldosterone synthesis and secretion' (Bast et al., 2018)) or stress (4926 'Relaxin signaling pathway') induced by LPS neuroinflammation. Substance abuse disorders which provoke neuroinflammation (5030 'Cocaine addiction', 5031 'Amphetamine addiction', 5032 'Morphine addiction', 5033 'Nicotine addiction', 5034 'Alcoholism' (Mitchell et al., 2019)) and pathways relating to repair (3440 'Homologous recombination', 3460 'Fanconi anemia pathway').

KEGG ID	Pathway Name	Notable Genes
4064	NF-kappa B signaling pathway	ICAM1
4080	Neuroactive ligand-receptor interaction	C3
4650	Natural killer cell mediated cytotoxicity	ICAM1
4668	TNF signaling pathway	ICAM1
5134	Legionellosis	C3
5142	Chagas disease (American trypanosomiasis)	C3
5143	African trypanosomiasis	ICAM1
5160	Hepatitis C	MX1
5162	Measles	MX1
5164	Influenza A	MX1, ICAM1
5165	Human papillomavirus infection	MX1
5167	Kaposi sarcoma-associated herpesvirus infection	C3, ICAM1
5168	Herpes simplex virus 1 infection	C3
5169	Epstein-Barr virus infection	ICAM1
5203	Viral carcinogenesis	C3

Table 4.5: **Enriched Pathways Containing Genes Up-regulated by Microglia**. Approximately 39% of pathways significantly enriched by PE-Aw for 3Scc-HA contain genes known to be up-regulated in astrocytes by the action of microglia. Whilst the majority are disease related, pathways 4064 and 4668 are known to play a role in response to LPS induced neuroinflammation with pathway 4080 containing proteins responsible for anti-inflammatory response. Notable genes taken from (Qiu et al., 2018).

Several signaling pathways whose role in microglial inflammation response were also enriched: 4012 'ErbB signaling pathway' (Xu et al., 2017), 4330 'Notch signaling' (Grandbarbe et al., 2007), 4371 'Apelin signaling pathway' (Chen et al., 2015). Additionally the enrichment of the 'Cortisol synthesis and secretion' pathway highlights the role of cortisol in inhibiting microglial activation after LPS stimulus (Drew and Chavis, 2000).

The significant enrichment of all synapse related pathways is notable and may reflect the non-cell autonomous role of microglia in synaptic stripping, which can occur in response to oxidative stimuli such as those resulting from LPS (Kettenmann et al., 2013), and neuroprotective synaptic migration, also demonstrable as a result of LPS stimulus (Chen et al., 2014).

For 3Scc-HA, we have 4 genes of interest from (Qiu et al., 2018) that are upregulated by microglia: markers of neurotoxic "A1" type astrocytes C3 and MX1 as well as CHI3L1 and ICAM1, the latter two being aberrantly expressed by astrocytes in a range of neurodegenerative diseases. As expected, pathways involving all these

genes, with the exception of CHI3L1, which is not annotated to any KEGG pathways, are significantly enriched as can be seen in Table 4.5. TNF signaling has also been observed as a response of astrocytes to LPS in other research (Sharif et al., 1993).

Interestingly the pathways 3010 'Ribosome', 4064 'NF-kappa B signaling pathway', 4620 'Toll-like receptor signaling pathway' and 4621 'NOD-like receptor signaling pathway' that were expected in the microglia dataset, due to their appearance in related research (Holtman et al., 2015), are strongly enriched for the astrocyte dataset. The 4060 'Cytokine-cytokine receptor interaction', that we may have expected given the differential expression of cytokines for the microglia dataset, is also significantly enriched here. Whilst toll-like receptor signaling is commonly associated with microglia, it has been observed in astrocytes in response to neuroinflammation (Alfonso-Loeches et al., 2010) particularly in presence of microglia (Holm et al., 2012).

As astrocytes themselves do not respond directly to LPS (Holm et al., 2012), we can expect the enrichments produced by the Pathway Entropy methods to relate to their interaction with microglia. The fact that we see the presence of many pathways we might expect in the microglia results for the astrocyte results is consistent with known interaction between microglia and astrocytes in response to LPS (Holm et al., 2012). Indeed the enrichment of the pathways 4060 'Cytokine-cytokine receptor interaction' and 4062 'Chemokine signaling pathway' may represent the interaction of the astrocytic cells with the respective proteins known to be released by microglia when stimulated by LPS.

With regards to an astrocytic response to neuroinflammation, we see several of the pathways identified in the previous chapter also enriched here: 3008 'Ribosome biogenesis in eukaryotes', 3010 'Ribosome' and 4630 'JAK-STAT signaling pathway'.

Lastly for 3Scc-MN we have five key genes of interest from (Qiu et al., 2018) among the many differentially expressed as a result of the LPS-induced microglia: Oasl2, Isg15, Ifitm3, Timp1 and Stat2. Of these 5 genes Ifitm3 does not belong to any pathway, Oasl2 and Timp1 are also annotated only to single pathways that are not enriched. However for Stat2 we see a wide enrichment of its annotated pathways, of which the following are of most interest: 4062 'Chemokine signaling pathway', 4217 'Necroptosis', 4380 'Osteoclast differentiation', 4621 'NOD-like receptor signaling pathway', 4630 'JAK-STAT signaling pathway'.

The neuron dataset also seems to find significant several pathways we expected to see (Holtman et al., 2015) in the microglia dataset: 3010 'Ribosome', 4145 'Phagosome', 4612 'Antigen processing and presentation', 4620 'Toll-like receptor signaling pathway' and 4621 'NOD-like receptor signaling pathway'. The phagosome pathway has been shown to be activated downstream in neurons by microglia after LPS stimulus (Bodea et al., 2014); in addition all but one (5322) significantly enriched KEGG

pathways identified by this study (4145, 4380, 4610, 5150, 5152, 5168) are found to be significant by PE-Aw. Whilst reference in the literature was found for the interaction of astrocytes and microglia for these pathways, no such reference was found to support this behaviour in neurons.

We do however observe significantly enriched pathways that we would expect to see: pathways relating to metabolism (10 'Glycolysis / Gluconeogenesis'), intercellular signaling (4080 'Neuroactive ligand-receptor interaction', 4514 'Cell adhesion molecules (CAMs)') and pathways relating to cell death as a result microglial LPS stimulation (Jeohn et al., 1998) (4217 'Necroptosis'). Indeed the TNF signaling pathway's (4668) involvement in neurodegeneration via LPS induced microglia is known (Jeohn et al., 1998). Chemokines are thought to be used by neurons to trigger a microglial response (Biber et al., 2008), thus the significant enrichment of the 4062 'Chemokine signaling pathway', in addition to other signaling pathways involving chemokines (4060 'Cytokine-cytokine receptor interaction', 4064 'NF-kappa B signaling pathway', 4620 'Toll-like receptor signaling pathway', 4621 'NOD-like receptor signaling pathway', 4657 'IL-17 signaling pathway'), is pertinent and might explain the unexpected presence of pathways 4620 and 4621 in particular.

It is worth recognising that enrichment of disease pathways accounts for just over 40% of the total enrichments, however the presence of the majority of these is a result of strongly expressed genes such as Stat2 and shared pathway components. Interestingly graft rejection pathways were observed in 3Scc-MN (5330 'Allograft rejection', 5332 'Graft-versus-host disease'). This could however simply be a result of shared genes, with both these pathways containing genes in the well represented TNF signaling pathway, in addition to 05332 containing representation of other chemokine related pathways (4060, 4620).

Figure 4.2: **Significantly Enriched,** $q < 0.1$ **Pathways for PE-Aw on the species separated 3Scc dataset.** The species separated datasets of 3Scc: human astrocyte (3Scc-HA), mouse neuron (3Scc-RN) and rat microglia (3Scc-RM) display distinct pathway enrichment profiles, uncovered by the application of PE-Aw. Pathways are labelled by name in the KEGG pathway database and the level of enrichment is reflected by a colour scale representing -log10(q-value).

4.4 DISCUSSION

Evaluation of SARGASSO Performance

For both of the datasets in this chapter, AD and 3Scc I used Sargasso to assign reads to their species of origin, producing datasets of size to be used in subsequent analysis. For the AD dataset, separated previously using the 'conservative' filtering strategy in Section 2.3.9.1, I used Sargasso's 'best' strategy to partition the dataset between mouse and rat with dataset sizes reflecting cell populations in the experimental cultures. Only a small amount of data here (2-3%) was lost due to similarity between the species genomes and the assignment levels observed reflected a higher population of rat neurons, which saw a mean of 78.3% of reads mapped to the genome assigned, over mouse, which saw a mean of 38.5% assigned. No problems were anticipated or observed here as Sargasso has been tested for separations between two species datasets of these species.

For the 3Scc dataset, in Section 4.3.2.1, I applied Sargasso to disambiguate RNA-Seq reads from three species: mouse neurons (3Scc-MN), human astrocytes (3Scc-HA) and rat microglia (3Scc-RM). As Sargasso had not been applied to a mixed-species dataset with three species this application demonstrated that it is able to separate such a dataset. As a result of the additional species, a higher number of reads were classed as ambiguous (with a mean of 4.1%, 6.6% and 5.0% of total mapped reads for mouse, human and rat respectively) due to the enlarged area of inseparable genome between the three species. The proportion of erroneous assignment to the rat for the neuron and astrocyte only conditions proved to be no higher than previously observed for the 'best' filtering strategy. Again the number of reads assigned reflected the relative cell numbers of the cell-types (with a mean of 79.4%, 21.5% and 16.8% of total mapped reads for mouse, human and rat respectively), as there are more neurons present than glia. Indeed the higher number of reads in 3Scc-MN is the likely cause of the subsequent pathway enrichment returning a higher number of results.

That the 3Scc dataset was comprised of longer reads, at 75bp compared to the 50bp of the other datasets, would have led me to expect a lower rate of mis-assignment, particularly given the result for Sargasso on simulated data of 150bp in Section 2.3.1. Seeing a higher percentage than expected would lead me to believe that sequencing a three species mixed-species RNA-Seq dataset at 50bp would be inadvisable. Given that we also see some fluctuation in read assignment, for all species in the second replicate of the 3Scc LPS condition and for microglia on the third replicate of the same condition, this does potentially raise questions about the consistency of

the co-cultured cells certainly with regard to cell numbers and potentially to purity. Unfortunately however, for both the 3Scc and AD datasets, matched-purity information was not available, with purity having only been measured as confirmation of successful co-culture.

The additional figures for the separation of protein coding reads by Sargasso for these mixed species datasets demonstrates that we do not see the worse performance we might expect for highly conserved reads. Indeed the proportion of protein coding reads lost closely resembles the proportion of all reads not assigned by Sargasso for each replicate, frequently even marginally lower. However we do observe more variation here for the 3Scc separation than for the AD, however this would be expected given the additional uncertainty.

As a result, whilst more caution should perhaps be taken with the interpretation of the downstream results for the 3Scc dataset, we have demonstrated the utility of Sargasso on two additional datasets and for its use with a three species mixed-species RNA-Seq dataset. The performance of Sargasso in this chapter demonstrates both that it has generaliseable utility for mixed-species RNA-Seq separation across different experimental types, and with differing numbers of species, and that the resultant separations are well suited for downstream analysis. Indeed the use of this tool on these datasets has enabled the subsequent investigation of non-cell-autonomous effects for the cell-types involved.

Evaluation of Pathway Entropy Performance

The application of PE-Aw to the AD dataset, in Section 4.3.1, produced significant pathway enrichments appropriate to the experimental context, in this case neuron and astrocyte response to neuron firing, for ADcc-MA. When we consider the enrichments produced by PE-Aw on the ADcc-MA dataset we see a strong representation of both pathways relating to astrocytic function and also to NCAE. Pathways relevant to NCAE between astrocytes and neurons were detected, particularly metabolic support and the astrocyte-neuron lactate shuttle. PE-Aw additionally illuminated astrocytic support to neurons in relation to the experimental stimuli by enriching pathways describing astrocytic processing of excess level of glutamate and GABA, which would be produced by neural firing provoked exposure to BiC and TBOA. These findings serve to demonstrate the capability of the PE-Aw method in providing pathway enrichment that is appropriate to the experimental context by incorporating a higher level of information.

Importantly the pathways enriched for ADcc-MA serve to confirm the experimental findings in (Hasel et al., 2017), reinforcing evidence for the NCAE role of astro-

cytes, through the astrocyte-neuron lactate shuttle, in providing metabolic support for neurons.

Whilst ADcc-RN produced a number of p-significant pathways, none of these enrichments passed multiple testing correction. Whilst the network produced a higher number of clusters than did ADcc-MA, 21 compared to 20, this additional complexity has not resulted in the strong clustering of pathway members that PE-Aw requires. It is perhaps of note that the AD dataset has a different experimental design, a control and two drug stimuli, from the OS (control, stimulus, post-stimulus) and 3Scc datasets (2 controls, 2 LPS stimulus). This results in a network influenced more by experimental stimulus than resting state. However, that the strong performance of PE-Aw on ADcc-MA is seemingly unaffected by any impact the design may have would preclude this conclusion, it is possible that increasing the number of permutations could have a positive impact here.

The results for the 3Scc dataset, in Section 4.3.2, can also be observed to produce enrichment results appropriate to the experimental context, that of LPS induced microglial interaction, and also of NCAE between the cell-types.

In the microglia, whilst we certainly see pathways relevant to inflammation response and cell-survival, we do not see several pathways involving Toll-like receptor 4 (TLR4) that we would expect from previous studies with this stimulus (NF-kappa B, Toll-like receptor and NOD-like receptor signaling pathways) (Holtman et al., 2015). Indeed further investigation showed the majority of these pathway's members to not be strongly co-expressed with one another, leading to them being assigned to different clusters. Given that differential expression analysis on the same data has shown genes involved in these pathways to be differentially expressed (Qiu et al., 2018), I think it likely that 2 conditions in triplicate is not enough to elicit a strong co-expression network with WGCNA, as I will discuss later, and so perhaps in this instance Pathway Entropy is not the most appropriate tool for analysis.

The astrocytes, 3Scc-HA, interestingly exhibit the much of the pathway response we would expect from the microglia, which has previously been observed in co-culture studies with LPS stimuli (Holm et al., 2012). Pathways relevant to the action of astrocytes in response to neuroinflammation were also significantly enriched.

The neuron dataset, 3Scc-MN, results include pathways previously observed to be enriched by neurons in neuron-microglia co-cultures stimulated with LPS (Bodea et al., 2014), in addition to pathways relating to neural function and synaptic activity. That TLR4 related signaling pathways are enriched here is interesting, as whilst neuron-micoglia chemokine signaling is hypothesised to take place (Biber et al., 2008), it has not yet been experimentally confirmed.

Non-cell autonomous effects can thus be seen to be well represented in the astrocyte and neuron datasets, with evidence for several observed and theorised intercellular processes present in the enrichments, the latter providing an avenue for further investigation. Whilst the application of PE-Aw to 3Scc-RM did not enrich the expected inflammation response pathways, it did enrich pathways relating to the NCAE role of microglia in stripping and neuro-protective migration of synapses (Kettenmann et al., 2013; Chen et al., 2014). We can therefore conclude that PE-Aw once again produces results appropriate to the data and importantly, that the number of species present in the source data does not prohibit such analysis.

With relation to prior investigation on the 3Scc data in (Qiu et al., 2018), we can see that differentially expressed genes of interest for each cell-type were represented in the significant pathway enrichments produced by PE-Aw. Whilst we cannot always expect this representation, as their very presence depends on their annotation in the database used, a manual task in the case of KEGG, these results provide additional validation for the results of this previous investigation.

The Impact of Sequencing Variables and Sample Quantity on Analysis

The two datasets analysed in this chapter (AD, 3Scc) and the OS dataset all have differences with regards to their sequencing, in addition to their experimentation and focus. Whilst the OS dataset consisted of three conditions in triplicate at a read depth of 50 million reads, where the conditions were effectively control-stimulus-post-stimulus, the datasets we used in this chapter differed in one or more respects. The AD dataset consisted of three conditions, control-stimulus A-stimulus B, with 4 replicates and at a read depth of 150 million reads. Lastly, the 3Scc dataset consisted of 4 conditions, control-stimulus both with and without microglia present, in triplicate at a read depth of 100 million reads for the conditions without microglia and 220 million in the conditions with.

The effect of the different experimental setups, with regard to the conditions, is particularly noticeable with the PE-Aw analysis of the AD dataset where we see well characterised enrichments for one species separated dataset and no significant enrichments for the other. The Sargasso separation results however would give no indication themselves of such an eventuality.

The OS dataset and the 3Scc dataset, for the mouse and human data, on the other hand demonstrate well characterised and appropriate enrichments. Whilst the 3Scc dataset doesn't have a post-stimulus condition as the OS dataset, it does have an additional condition for the mouse and human data that appears to aid characterisation of pathway enrichment. 3Scc-RM does suffer, conversely, from lack of data

as only two of the four conditions contain the microglia and these are only in triplicate, giving 6 samples to work with, which is lower than WGCNA's recommended minimum of 9. As a result we do not see representation of a number of pathways we would expect to see in microglia under LPS stimulation, however the significant enrichments produced do well portray the NCAE role of the cell-type and response to the experimental stimulus.

The sequencing depth will almost certainly play a role in influencing the results of the Pathway Entropy method, simply because building the co-expression networks with more data will result in better characterised co-expression between genes and thus aid subsequent clustering of the network. That we see the strongest enrichment results, in terms of quantity of enrichments, for the dataset containing the highest read depth is likely no coincidence.

5

GENERAL DISCUSSION

The research in this thesis aimed to increase our ability to study and analyse non-cell-autonomous effects (NCAE) between cell-types through the development and application of novel tools. Investigation was made into enabling the partition of gene expression data by cell type with greater accuracy and less potential risk of bias and spurious gene expression than existing techniques, in addition to the investigation of NCAE using methods from network biology and information theory. The product of this research has been two tools: Sargasso and Pathway Entropy.

Following the motivation for study and the introduction to the relevant background in Chapter 1, I described the design and development of Sargasso in Chapter 2. Here I evaluated its performance, in terms of assignment accuracy, read loss and false positives, on both simulated and experimental RNA-Seq datasets, in addition to identifying the effect of assignment variables and species choice on overall separation. I investigated the impact of Sargasso's use on downstream analysis and compared this with the effect of contamination frequently observed in physical separation. In Chapter 3 I detailed the design and implementation of the Pathway Entropy pipeline, which uses a combination of methods from network biology, information theory and draws upon previous research (Gambardella et al., 2013). I applied this method, using the KEGG pathway database, to experimental RNA-Seq data and evaluated the contribution of additional levels of information on the enrichments produced, and also investigated its potential to act as a means of evaluating network construction parameters. I compared the performance of this new method to both its predecessor (Gambardella et al., 2013) and to established methods for pathway enrichment analysis. Subsequently in Chapter 4 I applied my two novel methods to two different experimental mixed-species RNA-Seq datasets to both better evaluate their consistency of performance over additional data and to investigate the NCAE between cells in these datasets, indeed confirming the activity of pathways observed in previous studies.

Here I draw from the evidence presented in the thesis to summarise the contributions of these two novel methods before detailing their limitations and those present in the study. Lastly I will provide a summary of the key points for future work that I have already presented in Chapters 2 and 3.

5.1 CONTRIBUTION OF SARGASSO

In order to study NCAE we must be able to confidently attribute the information of study, in our case gene expression data derived from RNA-Seq, to the correct cell-types involved in an experiment. Previously this partitioning has been carried out primarily through the use of physical techniques, such as those introduced in Section 1.2.3.2. These techniques introduce bias to the derived gene expression data, through the activation of stress and cell death related genes, in addition to risking contamination by non-desired cell-types through imperfect separation, whose presence can have a sizeable impact on derived gene expression as I have discussed in Section 2.3.6.

In this project I have developed Sargasso which presents an *in silico* alternative to these approaches. Through the use of mixed-species RNA-Seq cultures where each cell-type belongs to a distinct species, Sargasso assigns reads individually based on the quality of their alignment to the genomes of the species involved thus circumventing the bias inherent in physical separation techniques. Sargasso demonstrates a high level of accuracy for this assignment and whilst a small number of false positives are introduced, through incorrect species assignment, the impact of its use on downstream analysis has been demonstrated to be minimal, far smaller than the impact of cell-type contamination. Sargasso has been demonstrated here to perform well in scenarios involving datasets containing both two and three distinct cell-types and its performance on solely protein coding reads, an area of potentially higher conservation, is comparable to that across all reads. Its use has enabled the investigation of NCAE on the applied datasets.

The ability to disambiguate RNA-Seq reads by cell type of origin *in silico*, at minimal risk of misinterpreting the data, presents a key step forward in our ability to accurately study NCAE. There are presently no other methods that can perform this disambiguation adequately on this type of data, with the closest *in silico* methods, designed for use in xenograft studies, failing to function for a wider range of species or producing a high level of mis-assignments. The resultant separations also provide higher resolution data than is presently possible using *in silico* deconvolution methodologies. As a result, Sargasso enables general purpose study of NCAE between cell-types that can be cultured *in vitro*. It is hoped that the availability of Sargasso and its peer-reviewed applications (Hasel et al., 2017; Qiu et al., 2018) will enable other research groups to further their pursuit of NCAE research.

5.2 CONTRIBUTION OF PATHWAY ENTROPY

Biological pathways represent our present understanding of the role of genes and other molecules in key biological processes from metabolism to immune response, and they therefore are important repositories of knowledge than can be drawn from to inform our understanding and provide context to gene expressions studies. Enrichment analyses that make use of these pathways commonly use no more information to inform their results than a set of gene identifiers, with which many use a variant of Fisher's exact test to determine the over-representation of any particular pathway's members in a set of genes of interest. Many of these methods have been in existence and continued use for many years, during which time the complexity of bioinformatics analysis and additional contextual information has increased. However whilst some new methods have been developed to better utilise such data, many long-standing methods have not optimised their usage of this information to inform their analysis.

In this thesis I have presented the development and application of the Pathway Entropy methodology, most notably PE-Aw, which is an alternative to existing pathway enrichment methods that is specifically designed for use with clustered gene co-expression networks, such as those produced by WGCNA. The Pathway Entropy methodology builds upon the DiNA tool to use information theoretic entropy to determine an enrichment based upon the representation of a pathway's genes within the network clusters. This method makes additional use of the correlation of expression between pathway gene members, the edge weights that underpin the network, to inform its enrichments. Pathway Entropy also normalises for pathway and cluster size in addition to using pathway specific null distributions for significance calculation.

The use of this additional information by PE-Aw has resulted in consistently relevant and appropriate enrichments when applied to experimental data outperforming the original DiNA approach in this regard, indeed uncovering evidence for NCAE in five out of six cell-type specific datasets and providing additional validation for the results of our analysis in (Hasel et al., 2017; Qiu et al., 2018). Its incorporation of pathway size normalisation addresses the concern of pathway bias by the DiNA methodology. In addition when tested against long used hypergeometric methods, unique enrichments by PE-Aw displayed evidence of biological action undetected by clusterProfiler. Our method's increased use of network specific information (network edges and edge weights) also serves to increase confidence in pathway enrichments, with both information sources having a demonstrable positive impact on enrichment results. As a result, Pathway Entropy presents the most appropriate method, of those

tested, to investigate pathway involvement in gene co-expression networks, whose use is becoming increasingly commonplace in bioinfomatics research. Its application here in investigating NCAE has also demonstrated its utility for this specific avenue of study. Whilst the investigation into its use as an evaluator for network method choice was curtailed by project time constraints, it does highlight cluster optimisation as a separate topic of future importance to the development of the tool.

The data produced in the course of the Pathway Entropy analysis is also of great value to further investigation into the NCAE between neurons and astrocytes or neurons, astrocytes and microglia, with the network and resultant clusters suitable for use with other network based tools or for more forensic insight with regard to Pathway Entropy's enrichments.

5.3 LIMITATIONS OF THIS PROJECT

5.3.1 *Limitations of the Source Data*

The source data used for this project is entirely neurological in focus, with the OS and AD datasets consisting of neurons and astrocytes and the 3Scc dataset containing microglia in addition. The rational for this was the investigation of NCAE between these specific neurological cell-types, and in this we have been successful (Hasel et al., 2017; Qiu et al., 2018). As a result of this however, whilst the novel methods presented here have been demonstrated to perform with consistency over these datasets, we cannot say with certainty that this performance is generalisable to data from other biological domains. However their performance as demonstrated in this thesis leaves us no reason to believe they will not be generalisable. Nevertheless, this thus presents an obvious opportunity for future work with both methods.

The culturing and sequencing also present limitations for the downstream analysis. The cell cultures do not contain an equal proportion of cells from either cell-type, as can be seen in the proportion of assigned reads particularly for the 3Scc dataset in Chapter 4, and matched-sample cell purity information was not available to validate these variations. Whilst the depth of sequencing, a factor limited by cost, should compensate for the difference in cell numbers, as has been intended in the sequencing used, it is possible that effects may still persist in the sequenced data. The higher depth of sequencing for the 3Scc dataset, particularly for 3Scc-MN, is likely a contributing factor to the increased sensitivity of the Pathway Entropy results. The length of reads presents another limitation as longer reads, particularly for Sargasso, are more likely to contain species specific information and therefore are more likely to be correctly assigned to their species of origin under the less stringent 'best' fil-

tering strategy, as we have demonstrated with simulated 150bp reads in Chapter 2. Longer read sequencing also presents an increased cost however, in addition to potential error risk dependent on sequencing technology, and as a result two of our datasets (OS and AD) have short 50bp reads with only the 3Scc dataset having longer 75bp reads.

The co-culturing of mixed-species cells also raises the issue of compatibility. Whilst we have tackled this in our paper (Hasel et al., 2017), other research we have carried out also demonstrates species specific differences in gene expression for function between mouse and human neurons which had previously not been investigated (Qiu et al., 2016). Therefore through the use of mixed-species cultures we will be introducing species specific differences in cellular function, even if minor, that should be taken into account when drawing conclusions about the data.

5.3.2 *Limitations of Network Methods*

Due to the time constraints of the project, I have used the WGCNA tool to construct the networks for analysis with Pathway Entropy using default parameters. Previous studies have identified inadequacies in WGCNA's clustering strategy (Botía et al., 2017) and proposed alternatives. However whilst I have not had sufficient time to implement and evaluate the effect of these alternative strategies on Pathway Entropy this presents a logical direction for future work. Similarly Pearson's correlation coefficient has been used due to it being WGCNA's default correlation method, however Spearman's rank correlation coefficient and biweight midcorrelation are unexplored and are possible alternatives.

Pathway Entropy's bootstrapping procedure requires the generation and export to file of two thousand gene co-expression networks. WGCNA uses a thresholding technique called 'soft-thresholding' which does not remove any edges between nodes; rather, it simply scales them using a soft thresholding power to emphasise stronger connections over weaker ones. An issue with this technique however is that it means exported network files are large as the edge lists include all possible pair-wise combinations of potentially tens of thousands of genes. This presents a computational storage issue, especially when working with multiple datasets for which each species requires bootstrapping and also a computational processing issue when loading these datasets to extract their contents. I have thus implemented a cut-off threshold of 0.2 correlation for the export of these networks for reasons of computational practicality, as below this point file size increases sharply by orders of magnitude. Whilst this cut-off is lower than that used in other studies working with gene co-expression networks (Ficklin et al., 2010; Ficklin and Feltus, 2011; Gobbi and Jurman, 2015; Wang

et al., 2014), it is removing data that might otherwise contribute to a more robust null distribution.

5.3.3 *Limitations of Pathway Entropy's Enrichment*

The enrichment methodology employed by Pathway Entropy sets a high bar for pathways to be enriched as it requires pathway members to be primarily co-expressed together and, as a result, clustered together. This form of enrichment therefore assumes that for a pathway to be active, we would have to see similar activity amongst all its members. Whilst for some small pathways this is a fair assumption, for larger or diverse pathways it may be possible that different portions of the pathway demonstrate different patterns of co-expression, perhaps in relation to their interaction with other genes, and are therefore clustered separately, resulting in a poor enrichment. An example of this could be 3Scc-RM. Previous research had led to the expectation of the enrichment of several pathways (particularly NOD-like and toll-like receptor signaling pathways and the NF-kappa B signaling pathway) (Holtman et al., 2015). The members of these pathways were present in the data and many were co-expressed, however these members were split across three or four separate clusters resulting in a poor entropy score and therefore low significance. Whilst its possible that additional data could have better informed the clustering, its also possible that the co-expression of these genes with other genes in the network would remain stronger than those of their other pathway members and thus obscure detection with this method.

A similar issue is the overlap of pathway members between pathways which can, as we have mentioned in Chapters 3 and 4, lead to the spurious enrichment of notionally unrelated pathways. This is hard to avoid with the methodology as it stands as many genes do have functionally important roles in different pathways, especially when using disease pathways, and this therefore presently requires users to cast a critical eye over the enrichment results. However such caution should not be unique to Pathway Entropy, indeed the results of enrichment tools should be always be appraised with a critical eye. This is something that could potentially be mitigated by controlling for genes unique to particular pathways, thereby only enriching a pathway if there is enough representation of unique members.

Pathway Entropy presently uses permutation testing to determine pathway significance. This means of testing is very computationally expensive and presents a constraint on the wider applicability of the pathway entropy methodology.

5.4 SUMMARY OF FUTURE RESEARCH DIRECTION

I have detailed intended future work in Chapters 2 and 3 for the Sargasso and Pathway Entropy methods respectively. Here I will summarise this in light of the additional application of these methods in Chapter 4.

5.4.1 *Sargasso*

In Chapter 4 I applied Sargasso to a mixed-species RNA-Seq dataset containing three species, demonstrating the tool's applicability beyond two species datasets. Its extension to four and five species separations have since been undertaken by colleagues in my research group with similar success, showing that beyond three species few additional reads are lost due to similarity as the species present cover the majority of highly conserved mammalian genes.

Future research with Sargasso could thus be directed towards working with laboratory mouse strains using the Sanger mouse genomes project. The ability to demonstrate separation in this area would lead to a more robust protocol for NCAE study, removing any doubt about species compatibility or of gene expression differences. In order to pursue this aim, and in an effort to refine the current protocol, the application of machine learning methods to the species assignment algorithm is a worthy avenue of exploration to increase separation accuracy. By this I mean primarily the reduction of false negatives under conservative assignment thresholds; however the reduction of false positive mis-assignments under less stringent assignment thresholds would be an additional boon.

The computational efficiency of Sargasso also presents an opportunity for further work. Pre-identification of highly conserved genomic areas, in correspondence to dataset read length, could reduce overall processing time by removing regions in which reads are inseparable. Introducing functionality for automated integration of separate genomic libraries, primarily those containing types of RNA not included in reference genomes as standard (e.g. rRNA), would also serve to increase accuracy by preventing rejection of reads in these domains.

5.4.2 *Pathway Entropy*

Application of the Pathway Entropy method in Chapter 4 was particularly illuminating for informing future research direction with the tool. The seeming dependence of sensitivity on dataset quality, given the increased quantity of enrichments produced

in the analysis of the 3Scc dataset, is particularly informative for future applications. Consistent throughout the applications of Pathway Entropy in this thesis is the spurious enrichment of pathways on the basis of shared genes, as discussed in Section 5.3.3. A key direction for future development of Pathway Entropy is to increase the accuracy of enrichment to mitigate this cross-talk of genes. This could be done through additional normalisation: either through normalising for the number of pathways co-expressed members are shared between or for nested representations of other pathways, with an enrichment requiring co-expression amongst each of its component parts. A threshold could also be trialled to require a minimum number of genes unique to a pathway to be co-expressed in order for an enrichment to be reported. The extension of the tool to work with Reactome would also introduce the possibility of weighting or normalising a pathway's enrichment based upon the co-expression of its hierarchical components in the network. As a different approach, enrichment of domain relevant ontologies, such as NiGO for neurological data, could be applied to significant enrichments and their top terms attached to the Pathway Entropy results for cross-reference.

The Pathway Entropy method as applied in this thesis has four variant implementations: PE-A and PE-T, both weighted and unweighted. Whilst PE-Aw was determined to be the better performing of the four, it is possible that a synthesis of the PE-A and PE-T approaches would allow for the benefits of increased information content without the restriction on usable pathway information; this could perhaps be done by adding a weighting for pathway members present in the topology over those who are not. This weighting could also contribute to a reduction in spurious enrichments as co-expressed genes in the pathway's topology are better indicators of pathway involvement than pathway members whose functional role is uncertain or unknown in relation to other members.

Whilst the accuracy of Pathway Entropy is of primary importance for future work, the generation of bootstrapping iterations for the permutation testing is presently a computationally intensive task both for processing and storage. More effective parallelisation or an alternative testing methodology would increase the practicality of the method's use.

Lastly a key consideration for future work in Pathway Entropy's methodology are the network methods that underpin it. Presently WGCNA is used as recommended, however research in this thesis into the minimum cluster size parameter indicates that cluster optimisation, a topic outside of the work in this thesis, is an important area to investigate to improve the methodology. This investigation should take into account not just the parameters easily alterable in WGCNA (e.g. correlation, minimum cluster size, merge distance) but also alternatives to the methods WGCNA implements,

particularly thresholding approach and clustering method, both of which have been areas of developmental focus in network biology in recent years.

5.5 CONCLUDING REMARKS

The scientific methods developed in this thesis have much potential for application beyond the scope of this research project.

The Sargasso tool now enables the study of NCAE with a higher degree of precision than previously available. Our publication of the protocol for mixed-species *in vitro* culturing and the use of Sargasso, in addition to the work in this thesis, will hopefully facilitate and inspire the future application of this methodology by other research groups in the pursuit of NCAE research.

The Pathway Entropy method at present, and with subsequent refining, presents a alternative to established enrichment techniques by utilising a higher content of information in its enrichments. Whilst applied here to NCAE, it is publicly available and can be applied to any gene co-expression network.

APPENDIX

A

SOURCE CODE, DATASET AVAILABILITY & PROJECT RESOURCES

A.1 PROJECT CODE

The source code for both the Sargasso and Pathway Entropy methods can be found on the electronic media submitted with the thesis, Table A.1 details the locations on the file-system within this media. This encompasses all code used for this thesis, including for the visualisation of results. The Sargasso code provided has been cloned from our public GitHub (`https://github.com/statbio/Sargasso`) on 2017-8-24 and has been provided as it was the exact version used to separate the AD and 3Scc datasets in Chapters 2 and 4, the most recent version of the Sargasso code can of course be found at this URL.

Whilst Sargasso is simple and user friendly to run, the Pathway Entropy scripts are still in a developmental stage and as such are not user friendly nor presently automated for pipeline execution.

All Python packages and dependencies for Sargasso can be installed using the provided script. All R packages for the Pathway Entropy code are publicly available on CRAN or Bioconductor.

Tool/Task	Chapter Used	Location
Sargasso	2, 4	`Appendices/Code/Sargasso/Sargasso_git_clone/`
Sargasso Visualisations	2	`Appendices/Code/Sargasso/Sargasso_plotting_scripts/`
Read Error Simulation Script	2	`Appendices/Code/Sargasso/Simulated_error_script/`
Pathway Entropy	3, 4	`Appendices/Code/Pathway_Entropy/Pathway_Entropy_source_code/`
Pathway Entropy Visualisations	3, 4	`Appendices/Code/Pathway_Entropy/Pathway_Entropy_plotting_scripts/`

Table A.1: **Location of Source Code**. This table summarises the locations of the source code used in this thesis.

A.2 EXPERIMENTAL DATA AVAILABILITY

The availability of the experimental datasets used in this thesis are listed, by chapter of use, in Table A.2. The samples I have used are also listed as the full datasets contain more samples than these.

Dataset	Chapter	Availability	Samples Used
Oxidative Stress (OS)	2, 3	Unpublished, Available on Request	1A1, 1A2, 1A3, 2A1, 2A2, 2A3, 3A1, 3A2, 3A3, 1N1
Activity Dependence (AD)	2, 4	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5514/	1F2, 2F2, 3F2, 4F2, 1F4, 2F4, 3F4, 4F4, 1F5, 2F5, 3F5, 4F5, RatOnly, A1, A2, A3, S1, S2, S3, AN1, AN2, BN1, BN2, CN1, CN2
Three Species (3Scc)	4	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5987/	J4NAC, K4NAC, L4NAC, J6NAL500, K6NAL500, L6NAL500, J7NAMC, K7NAMC, L7NAMC, J9NAML500, K9NAML500, L9NAML500

Table A.2: **Availability of Experimental Data**. This table summarises the availability of the experimental data used in this thesis.

A.3 RAW RESULTS DATA

Raw data generated by Sargasso, topGO and Pathway Entropy from their application in this thesis can be found in the electronic media submitted with this thesis, Table A.3 details their location on the file-system within this media. For Sargasso this consists of assignment summaries for the core datasets, for topGO this consists of Gene Ontology enrichment results and for Pathway Entropy this consists of the enrichment results produced for these datasets. Enrichments for DiNA, KEGGprofile and clusterProfiler have also been included. Enrichments include all pathways unconstrained by p-value.

The results for each tool are archived by dataset at the locations specified in Table A.3, where [Species] or [Dataset] refer to corresponding directory names.

A note on the codes in the filenames: MC-10 refers to minimum cluster size 10, MC-30 to minimum cluster size 30, 'Less-Sensitive' to significance testing using null-

distributions from permutations using MC-10 and 'Unsensitive' to significance testing using null-distributions from permutations using MC-30.

Tool	Data Type	Location
Sargasso	Simulated Read Assignment	`Appendices/Raw_Data/Sargasso_Assignment_Summaries/Simulated_Data/[Species]`
Sargasso	Read Assignment	`Appendices/Raw_Data/Sargasso_Assignment_Summaries/Experimental_Data/[Dataset]`
topGO	Gene Ontology Enrichment	`Appendices/Raw_Data/topGO_Enrichments/`
Pathway Entropy, DiNA, KEGGprofile, clusterProfiler	Pathway Enrichment	`Appendices/Raw_Data/Pathway_Enrichments/[Dataset]`

Table A.3: **Location of Raw Results Data**. This table summarises the locations of the raw results data produced by the work in this thesis.

A.4 FULL SIZE FIGURES

Full size versions of all figures included in this thesis can be found in the electronic media submitted with the thesis, Table A.4 details their location on the file-system within this media.

Chapter	Location
Chapter 1	`Appendices/Figures/Chapter_1_Introduction/`
Chapter 2	`Appendices/Figures/Chapter_2_Sargasso/`
Chapter 3	`Appendices/Figures/Chapter_3_Pathway_Entropy/`
Chapter 4	`Appendices/Figures/Chapter_4_Application_Chapter/`

Table A.4: **Location of Full Size Figures**. This table summarises the locations of the full size figure files included in the chapters of this thesis.

B

SUPPLEMENTARY INFORMATION

B.1 SUPPLEMENTARY TABLES

GO.ID	Term	Annotated	Significant	Expected	weight_fisher
GO:0007608	sensory perception of smell	1067	268	54.67	< 1e-30
GO:0007186	G-protein coupled receptor signaling pat...	1830	290	93.76	< 1e-30
GO:0050907	detection of chemical stimulus involved ...	295	44	15.11	9.0e-14
GO:0030216	keratinocyte differentiation	134	19	6.87	7.6e-07
GO:0018149	peptide cross-linking	58	14	2.97	9.6e-07
GO:0009952	anterior/posterior pattern specification	223	19	11.43	6.2e-05
GO:0048668	collateral sprouting	30	5	1.54	0.00041
GO:0007017	microtubule-based process	664	24	34.02	0.00071
GO:0060968	regulation of gene silencing	68	10	3.48	0.00167
GO:1900004	negative regulation of serine-type endop...	12	4	0.61	0.00243
GO:0018003	peptidyl-lysine N6-acetylation	2	2	0.1	0.00262
GO:0007586	digestion	126	7	6.46	0.00313
GO:0006446	regulation of translational initiation	56	10	2.87	0.00333
GO:0060017	parathyroid gland development	7	3	0.36	0.00402
GO:0048704	embryonic skeletal system morphogenesis	102	12	5.23	0.00503
GO:1990830	cellular response to leukemia inhibitory...	128	14	6.56	0.00597
GO:0007606	sensory perception of chemical stimulus	1377	291	70.55	0.00625
GO:0071630	nuclear protein quality control by the u...	3	2	0.15	0.00760
GO:0010482	regulation of epidermal cell division	3	2	0.15	0.00760
GO:0090135	actin filament branching	3	2	0.15	0.00760
GO:0030219	megakaryocyte differentiation	49	4	2.51	0.00760
GO:0051967	negative regulation of synaptic transmis...	17	4	0.87	0.00954
GO:0007004	telomere maintenance via telomerase	58	4	2.97	0.01232
GO:0006334	nucleosome assembly	103	15	5.28	0.01432
GO:1901386	negative regulation of voltage-gated cal...	19	4	0.97	0.01433
GO:0006342	chromatin silencing	101	12	5.17	0.01455
GO:2001206	positive regulation of osteoclast develo...	4	2	0.2	0.01468
GO:0046952	ketone body catabolic process	4	2	0.2	0.01468
GO:0006357	regulation of transcription from RNA pol...	1839	100	94.22	0.01522
GO:0071243	cellular response to arsenic-containing ...	11	3	0.56	0.01624

Table B.1: **Significantly Enriched GO Terms for Sargasso Lost Reads, 'conservative' Strategy.** This table lists the significant GO terms enriched through the application of topGO to the reads lost (rejected and ambiguous) by Sargasso using the 'conservative' filtering strategy. The top 30 most significant terms here have been displayed for brevity. Enrichment significance is measured using a weighted Fisher's exact test and is not multiple test corrected due to non-indepence of testing methodology as described in the tool author's paper: (Alexa and Rahnenführer, 2018)

GO.ID	Term	Annotated	Significant	Expected	weight_fisher
GO:0000122	negative regulation of transcription fro...	800	10	2.77	0.00041
GO:0071243	cellular response to arsenic-containing ...	11	2	0.04	0.00064
GO:0034063	stress granule assembly	12	2	0.04	0.00076
GO:0036120	cellular response to platelet-derived gr...	16	2	0.06	0.00137
GO:0045948	positive regulation of translational ini...	20	2	0.07	0.00216
GO:0030539	male genitalia development	23	2	0.08	0.00285
GO:1901389	negative regulation of transforming grow...	1	1	0	0.00346
GO:0071765	nuclear inner membrane organization	1	1	0	0.00346
GO:1905926	positive regulation of invadopodium asse...	1	1	0	0.00346
GO:1903743	negative regulation of anterograde synap...	1	1	0	0.00346
GO:1902684	negative regulation of receptor localiza...	1	1	0	0.00346
GO:0090134	cell migration involved in mesendoderm m...	1	1	0	0.00346
GO:0045234	protein palmitoleylation	1	1	0	0.00346
GO:1904840	positive regulation of male germ-line st...	1	1	0	0.00346
GO:0045944	positive regulation of transcription fro...	1118	10	3.87	0.00494
GO:0021543	pallium development	146	3	0.51	0.00676
GO:1904627	response to phorbol 13-acetate 12-myrist...	8	2	0.03	0.00682
GO:1905426	positive regulation of Wnt-mediated midb...	2	1	0.01	0.00691
GO:0010609	mRNA localization resulting in posttrans...	2	1	0.01	0.00691
GO:0043418	homocysteine catabolic process	2	1	0.01	0.00691
GO:0006850	mitochondrial pyruvate transmembrane tra...	2	1	0.01	0.00691
GO:2000814	positive regulation of barbed-end actin ...	2	1	0.01	0.00691
GO:0072752	cellular response to rapamycin	2	1	0.01	0.00691
GO:2000393	negative regulation of lamellipodium mor...	2	1	0.01	0.00691
GO:0045204	MAPK export from nucleus	2	1	0.01	0.00691
GO:1902164	positive regulation of DNA damage respon...	2	1	0.01	0.00691
GO:0050973	detection of mechanical stimulus involve...	2	1	0.01	0.00691
GO:0000381	"regulation of alternative mRNA splicing, ..."	38	2	0.13	0.00766
GO:0006355	"regulation of transcription, DNA-templat..."	3105	26	10.75	0.00966
GO:0010603	regulation of cytoplasmic mRNA processin...	8	2	0.03	0.01021

Table B.2: **Significantly Enriched GO Terms for Sargasso Lost Reads, 'best' Strategy.** This table lists the significant GO terms enriched through the application of topGO to the reads lost (rejected and ambiguous) by Sargasso using the 'best' filtering strategy. The top 30 most significant terms here have been displayed for brevity. Enrichment significance is measured using a weighted Fisher's exact test and is not multiple test corrected due to non-indepence of testing methodology as described in the tool author's paper: (Alexa and Rahnenführer, 2018)

BIBLIOGRAPHY

Aagaard, K., Riehle, K., Ma, J., Segata, N., Mistretta, T.-A., Coarfa, C., Raza, S., Rosenbaum, S., Van den Veyver, I., Milosavljevic, A., et al. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PloS one*, 7(6):e36466. (Cited on page 10.)

Adams, D. J., Doran, A. G., Lilue, J., and Keane, T. M. (2015). The mouse genomes project: a repository of inbred laboratory mouse strain genomes. *Mammalian Genome*, 26(9-10):403–412. (Cited on page 86.)

Ader, M. and Tanaka, E. M. (2014). Modeling human development in 3d culture. *Current opinion in cell biology*, 31:23–28. (Cited on page 13.)

Aggarwal, A., Guo, D. L., Hoshida, Y., Yuen, S. T., Chu, K.-M., So, S., Boussioutas, A., Chen, X., Bowtell, D., Aburatani, H., et al. (2006). Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer research*, 66(1):232–241. (Cited on page 102.)

Ahdesmäki, M. J., Gray, S. R., Johnson, J. H., and Lai, Z. (2016). Disambiguate: an open-source application for disambiguating two species in next generation sequencing data from grafted samples. *F1000Research*, 5. (Cited on pages 11, 75, 77, and 83.)

Alexa, A. and Rahnenfuhrer, J. (2010). topgo: enrichment analysis for gene ontology. *R package version*, 2(0). (Cited on page 19.)

Alexa, A. and Rahnenführer, J. (2018). Gene set enrichment analysis with topgo. `https://www.bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf`, Accessed: 2019-2-19. (Cited on pages 53, 195, and 196.)

Alfonso-Loeches, S., Pascual-Lucas, M., Blanco, A. M., Sanchez-Vera, I., and Guerri, C. (2010). Pivotal role of tlr4 receptors in alcohol-induced neuroinflammation and brain damage. *Journal of Neuroscience*, 30(24):8285–8295. (Cited on page 171.)

Allen, N. J., Bennett, M. L., Foo, L. C., Wang, G. X., Chakraborty, C., Smith, S. J., and Barres, B. A. (2012). Astrocyte glypicans 4 and 6 promote formation of excitatory synapses via glua1 ampa receptors. *Nature*, 486(7403):410. (Cited on page 14.)

Amann, A., Zwierzina, M., Gamerith, G., Bitsche, M., Huber, J. M., Vogel, G. F., Blumer, M., Koeck, S., Pechriggl, E. J., Kelm, J. M., et al. (2014). Development of an innovative 3d cell culture system to study tumour–stroma interactions in non-small cell lung cancer cells. *PloS one*, 9(3):e92511. (Cited on page 7.)

Anders, S., Pyl, P. T., and Huber, W. (2015). Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169. (Cited on page 38.)

Anderson, M. A., Burda, J. E., Ren, Y., Ao, Y., O'Shea, T. M., Kawaguchi, R., Coppola, G., Khakh, B. S., Deming, T. J., and Sofroniew, M. V. (2016). Astrocyte scar formation aids central nervous system axon regeneration. *Nature*, 532(7598):195. (Cited on page 147.)

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25. (Cited on page 18.)

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979. (Cited on page 12.)

Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M., Gingras, M.-C., Miller, D. K., Christ, A. N., Bruxner, T. J., Quinn, M. C., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47. (Cited on page 103.)

Ballouz, S., Dobin, A., Gingeras, T. R., and Gillis, J. (2018). The fractured landscape of rna-seq alignment: the default in our stars. *Nucleic acids research*, 46(10):5125–5138. (Cited on page 36.)

Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *science*, 325(5939):412–413. (Cited on page 102.)

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512. (Cited on page 102.)

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., and Grant, G. R. (2017). Simulation-based comprehensive benchmarking of rna-seq aligners. *Nature methods*, 14(2):135. (Cited on page 36.)

Bast, B.-O., Rickert, U., Schneppenheim, J., Cossais, F., Wilms, H., Arnold, P., and Lucius, R. (2018). Aldosterone exerts anti-inflammatory effects on lps stimulated microglia. *Heliyon*, 4(10):e00826. (Cited on page 169.)

Beisser, D., Klau, G. W., Dandekar, T., Müller, T., and Dittrich, M. T. (2010). Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics*, 26(8):1129–1130. (Cited on page 15.)

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325. (Cited on page 85.)

Bélanger, M., Allaman, I., and Magistretti, P. J. (2011). Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell metabolism*, 14(6):724–738. (Cited on page 147.)

Bellesi, M., de Vivo, L., Chini, M., Gilli, F., Tononi, G., and Cirelli, C. (2017). Sleep loss promotes astrocytic phagocytosis and microglial activation in mouse cerebral cortex. *Journal of Neuroscience*, 37(21):5263–5273. (Cited on page 163.)

Biber, K., Vinet, J., and Boddeke, H. (2008). Neuron-microglia signaling: chemokines as versatile messengers. *Journal of neuroimmunology*, 198(1-2):69–74. (Cited on pages 172 and 176.)

Bodea, L.-G., Wang, Y., Linnartz-Gerlach, B., Kopatz, J., Sinkkonen, L., Musgrove, R., Kaoma, T., Muller, A., Vallar, L., Di Monte, D. A., et al. (2014). Neurodegeneration by activation of the microglial complement–phagosome pathway. *Journal of Neuroscience*, 34(25):8546–8556. (Cited on pages 171 and 176.)

Botía, J. A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., Hardy, J., Lewis, C. M., Ryten, M., and Weale, M. E. (2017). An additional k-means clustering step improves the biological features of wgcna gene co-expression networks. *BMC systems biology*,

11(1):47. (Cited on pages 150 and 183.)

Bubela, T., Li, M. D., Hafez, M., Bieber, M., and Atkins, H. (2012). Is belief larger than fact: expectations, optimism and reality for translational stem cell research. *BMC medicine*, 10(1):133. (Cited on page 13.)

Burgess, D. J. (2019). Spatial transcriptomics coming of age. *Nature Reviews Genetics*, page 1. (Cited on page 8.)

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing, Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A., et al. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience*, 28(1):264–278. (Cited on pages 8 and 73.)

Callaham, M. L., Wears, R. L., Weber, E. J., Barton, C., and Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *Jama*, 280(3):254–257. (Cited on page 19.)

Chen, L., Tao, Y., and Jiang, Y. (2015). Apelin activates the expression of inflammatory cytokines in microglial bv2 cells via pi-3k/akt and mek/erk pathways. *Science China Life Sciences*, 58(6):531–540. (Cited on page 170.)

Chen, Z., Jalabi, W., Hu, W., Park, H.-J., Gale, J. T., Kidd, G. J., Bernatowicz, R., Gossman, Z. C., Chen, J. T., Dutta, R., et al. (2014). Microglial displacement of inhibitory synapses provides neuroprotection in the adult brain. *Nature communications*, 5:4486. (Cited on pages 170 and 177.)

Chintala, S. K., Fueyo, J., Gomez-Manzano, C., Venkaiah, B., Bjerkvig, R., Yung, W., Sawaya, R., Kyritsis, A. P., and Rao, J. S. (1997). Adenovirus-mediated p16/cdkn2 gene transfer suppresses glioma invasion in vitro. *Oncogene*, 15(17):2049–2057. (Cited on page 7.)

Christopherson, K. S., Ullian, E. M., Stokes, C. C., Mullowney, C. E., Hell, J. W., Agah, A., Lawler, J., Mosher, D. F., Bornstein, P., and Barres, B. A. (2005). Thrombospondins are astrocyte-secreted proteins that promote cns synaptogenesis. *Cell*, 120(3):421–433. (Cited on page 6.)

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630. (Cited on page 15.)

Chun, Y. S., Byun, K., and Lee, B. (2011). Induced pluripotent stem cells and personalized medicine: current progress and future perspectives. *Anatomy & cell biology*, 44(4):245–255. (Cited on page 13.)

Clarke, L. E. and Barres, B. A. (2013). Emerging roles of astrocytes in neural circuit development. *Nature Reviews Neuroscience*, 14(5):311. (Cited on page 14.)

Clarridge, J. E. (2004). Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, 17(4):840–862. (Cited on page 10.)

Clause, K. C., Liu, L. J., and Tobita, K. (2010). Directed stem cell differentiation: the role of physical forces. *Cell communication & adhesion*, 17(2):48–54. (Cited on page 13.)

Colombaioni, L. and Garcia-Gil, M. (2004). Sphingolipid metabolites in neural signalling and function. *Brain Research Reviews*, 46(3):328–355. (Cited on page 134.)

Colombo, E. and Farina, C. (2016). Astrocytes: key regulators of neuroinflammation. *Trends in immunology*, 37(9):608–620. (Cited on page 148.)

Conway, T., Wazny, J., Bromage, A., Tymms, M., Sooraj, D., Williams, E. D., and Beresford-Smith, B. (2012). Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics*, 28(12):i172–i178. (Cited on page 11.)

Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697. (Cited on page 91.)

Daverey, A., Drain, A., Crone, K., and Kidambi, S. (2014). Engineering of highly controlled in vitro co-culture model to study the mesenchymal stem cells mediated signaling in breast cancer cells. *Cancer Research*, 74(19 Supplement):3303–3303. (Cited on page 7.)

De Luca, A. C., Faroni, A., and Reid, A. J. (2015). Dorsal root ganglia neurons and differentiated adipose-derived stem cells: an in vitro co-culture model to study peripheral nerve regeneration. *JoVE (Journal of Visualized Experiments)*, (96):e52543–e52543. (Cited on pages 6 and 7.)

Denef, C. (2014). Contact-dependent signaling. *Cell Communication Insights*, 2014(6):1–11. (Cited on page 14.)

Devine, M. J., Ryten, M., Vodicka, P., Thomson, A. J., Burdon, T., Houlden, H., Cavaleri, F., Nagano, M., Drummond, N. J., Taanman, J.-W., et al. (2011). Parkinson's disease induced pluripotent stem cells with triplication of the α-synuclein locus. *Nature communications*, 2:440. (Cited on page 13.)

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21. (Cited on page 31.)

Drew, P. D. and Chavis, J. A. (2000). Inhibition of microglial cell activation by cortisol. *Brain research bulletin*, 52(5):391–396. (Cited on page 170.)

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210. (Cited on page 40.)

EMBL-EBI (2017). Illumina sequencing. `https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/illumina-`, Accessed: 2018-8-27. (Cited on pages 2 and 4.)

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T. J., Harrow, J., Guigó, R., et al. (2013). Systematic evaluation of spliced alignment programs for rna-seq data. *Nature methods*, 10(12):1185–1191. (Cited on page 36.)

Espina, V., Wulfkuhle, J. D., Calvert, V. S., VanMeter, A., Zhou, W., Coukos, G., Geho, D. H., Petricoin III, E. F., and Liotta, L. A. (2006). Laser-capture microdissection. *Nature protocols*, 1(2):586. (Cited on pages 7 and 9.)

Faisal, F. E., Zhao, H., and Milenković, T. (2015). Global network alignment in the context of aging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(1):40–52. (Cited on page 16.)

Fiannaca, A., La Paglia, L., La Rosa, M., Renda, G., Rizzo, R., Gaglio, S., Urso, A., et al. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC bioinformatics*, 19(7):198. (Cited on page 87.)

Ficklin, S. P. and Feltus, F. A. (2011). Gene co-expression network alignment and conservation of gene modules between two grass species: Maize (zea mays) and rice (oryza sativa). *Plant Physiology*, pages pp–111. (Cited on page 183.)

Ficklin, S. P., Luo, F., and Feltus, F. A. (2010). The association of multiple interacting genes with specific phenotypes in rice (oryza sativa) using gene co-expression networks. *Plant physiology*, pages pp–110. (Cited on page 183.)

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., and Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME journal*, 6(5):1007–1017. (Cited on page 10.)

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G., and Furlanello, C. (2018). Phylogenetic convolutional neural networks in metagenomics. *BMC bioinformatics*, 19(2):49. (Cited on page 87.)

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., Jovanovich, S. B., Nelson, J. R., Schloss, J. A., Schwartz, D. C., et al. (2009). The challenges of sequencing by synthesis. *Nature biotechnology*, 27(11):1013. (Cited on pages 33 and 37.)

Gambardella, G., Moretti, M. N., De Cegli, R., Cardone, L., Peron, A., and Di Bernardo, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics*, 29(14):1776–1785. (Cited on pages xiv, 3, 20, 25, 90, 92, 111, 112, 114, 119, 145, and 179.)

Geifman, N., Monsonego, A., and Rubin, E. (2010). The neural/immune gene ontology: clipping the gene ontology for neurological and immunological systems. *BMC bioinformatics*, 11(1):458. (Cited on page 152.)

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80. (Cited on page 91.)

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., Nair, V. S., Xu, Y., Khuong, A., Hoang, C. D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nature medicine*, 21(8):938. (Cited on page 12.)

Ghasemzadeh, A., ter Haar, M. M., Shams-bakhsh, M., Pirovano, W., and Pantaleo, V. (2018). Shannon entropy to evaluate substitution rate variation among viral nucleotide positions in datasets of viral sirnas. In *Viral Metagenomics*, pages 187–195. Springer. (Cited on page 20.)

Gligorijević, V., Janjić, V., and Pržulj, N. (2014). Integration of molecular network data reconstructs gene ontology. *Bioinformatics*, 30(17):i594–i600. (Cited on page 15.)

Gobbi, A. and Jurman, G. (2015). A null model for pearson coexpression networks. *PloS one*, 10(6):e0128115. (Cited on page 183.)

Gong, X., Hu, H., Qiao, Y., Xu, P., Yang, M., Dang, R., Han, W., Guo, Y., Chen, D., and Jiang, P. (2019). The involvement of renin-angiotensin system in lipopolysaccharide-

induced behavioral changes, neuroinflammation and disturbed insulin signaling. *Frontiers in pharmacology*, 10:318. (Cited on page 169.)

Grandbarbe, L., Michelucci, A., Heurtaux, T., Hemmer, K., Morga, E., and Heuschling, P. (2007). Notch signaling modulates the activation of microglial cells. *Glia*, 55(15):1519–1530. (Cited on page 170.)

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083. (Cited on page 40.)

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). Dingo: Differential network analysis in genomics. *Bioinformatics*, page btv406. (Cited on page 17.)

Halliwell, B. (2006). Oxidative stress and neurodegeneration: where are we now? *Journal of neurochemistry*, 97(6):1634–1658. (Cited on page 20.)

Hardingham, G. E. and Do, K. Q. (2016). Linking early-life nmdar hypofunction and oxidative stress in schizophrenia pathogenesis. *Nature Reviews Neuroscience*, 17(2):125. (Cited on page 20.)

Hasel, P., Dando, O., Jiwaji, Z., Baxter, P., Todd, A. C., Heron, S., Márkus, N. M., McQueen, J., Hampton, D. W., Torvell, M., et al. (2017). Neurons and neuronal activity control gene expression in astrocytes to regulate their development and metabolism. *Nature Communications*, 8:15132. (Cited on pages 8, 21, 25, 27, 28, 52, 59, 71, 72, 73, 74, 79, 80, 81, 82, 155, 162, 164, 175, 180, 181, 182, and 183.)

Hayes, W., Sun, K., and Pržulj, N. (2013). Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491. (Cited on page 16.)

Haynes, W. A., Tomczak, A., and Khatri, P. (2018). Gene annotation bias impedes biomedical research. *Scientific reports*, 8(1):1362. (Cited on page 19.)

Heiman, M., Kulicke, R., Fenster, R. J., Greengard, P., and Heintz, N. (2014). Cell type–specific mrna purification by translating ribosome affinity purification (trap). *Nature protocols*, 9(6):1282. (Cited on page 8.)

Hempel, C. M., Sugino, K., and Nelson, S. B. (2007). A manual method for the purification of fluorescently labeled neurons from the mammalian brain. *Nature protocols*, 2(11):2924. (Cited on page 8.)

Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L. A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from stanford. *Clinical chemistry*, 48(10):1819–1827. (Cited on page 8.)

Hoarau, J.-J., Krejbich-Trotot, P., Jaffar-Bandjee, M.-C., Das, T., Thon-Hon, G.-V., Kumar, S., W Neal, J., and Gasque, P. (2011). Activation and control of cns innate immune responses in health and diseases: a balancing act finely tuned by neuroimmune regulators (nireg). *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 10(1):25–43. (Cited on page 14.)

Holm, T. H., Draeby, D., and Owens, T. (2012). Microglia are required for astroglial toll-like receptor 4 response and for optimal tlr2 and tlr3 response. *Glia*, 60(4):630–638. (Cited on pages 171 and 176.)

Holtman, I. R., Raj, D. D., Miller, J. A., Schaafsma, W., Yin, Z., Brouwer, N., Wes, P. D., Möller, T., Orre, M., Kamphuis, W., et al. (2015). Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a

co-expression meta-analysis. *Acta neuropathologica communications*, 3(1):31. (Cited on pages 169, 171, 176, and 184.)

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13. (Cited on page 18.)

Hudson, N. J., Reverter, A., and Dalrymple, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput Biol*, 5(5):e1000382. (Cited on page 16.)

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular systems biology*, 8(1):565. (Cited on page 18.)

Ivanov, D. P., Parker, T. L., Walker, D. A., Alexander, C., Ashford, M. B., Gellert, P. R., and Garnett, M. C. (2015). In vitro co-culture model of medulloblastoma and human neural stem cells for drug delivery assessment. *Journal of biotechnology*, 205:3–13. (Cited on pages 6 and 7.)

Jensen, C. J., Massie, A., and De Keyser, J. (2013). Immune players in the cns: the astrocyte. *Journal of Neuroimmune Pharmacology*, 8(4):824–839. (Cited on page 162.)

Jeohn, G.-H., Kong, L.-Y., Wilson, B., Hudson, P., and Hong, J.-S. (1998). Synergistic neurotoxic effects of combined treatments with cytokines in murine primary mixed neuron/glia cultures. *Journal of neuroimmunology*, 85(1):1–10. (Cited on page 172.)

Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., and Wei, C. (2013). Nessm: a next-generation sequencing simulator for metagenomics. *PLoS One*, 8(10):e75448. (Cited on page 40.)

Johnson, M. A., Weick, J. P., Pearce, R. A., and Zhang, S.-C. (2007). Functional neural development from human embryonic stem cells: accelerated synaptic activity via astrocyte coculture. *The Journal of neuroscience*, 27(12):3069–3077. (Cited on pages 6 and 28.)

Kettenmann, H., Kirchhoff, F., and Verkhratsky, A. (2013). Microglia: new roles for the synaptic stripper. *Neuron*, 77(1):10–18. (Cited on pages 170 and 177.)

Kuchaiev, O. and Pržulj, N. (2011). Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396. (Cited on page 16.)

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., and Ma'ayan, A. (2018). Massive mining of publicly available rna-seq data from human and mouse. *Nature communications*, 9(1):1366. (Cited on page 36.)

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1. (Cited on pages xvi, 17, and 89.)

Langfelder, P. and Horvath, S. (2018). Wgcna manual. `https://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf`, Accessed: 2018-7-1. (Cited on page 104.)

Langfelder, P., Zhang, B., and Horvath, S. (2007). Branchcutting supplement. `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/BranchCutting/Supplement-published.pdf`, Accessed: 2018-7-1. (Cited on page 104.)

Lee, M., McGeer, E. G., and McGeer, P. L. (2011). Mechanisms of gaba release from human astrocytes. *Glia*, 59(11):1600–1611. (Cited on page 147.)

Li, Q. and Wang, Z. (2013). Influence of mesenchymal stem cells with endothelial progenitor cells in co-culture on osteogenesis and angiogenesis: an in vitro study. *Archives of medical research*, 44(7):504–513. (Cited on page 7.)

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930. (Cited on pages 34, 35, and 38.)

Liddelow, S. A., Guttenplan, K. A., Clarke, L. E., Bennett, F. C., Bohlen, C. J., Schirmer, L., Bennett, M. L., Münch, A. E., Chung, W.-S., Peterson, T. C., et al. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, 541(7638):481. (Cited on page 14.)

Linnartz, B., Bodea, L.-G., and Neumann, H. (2012). Microglial carbohydrate-binding receptors for neural repair. *Cell and tissue research*, 349(1):215–227. (Cited on page 169.)

Liska, A., Galbusera, A., Schwarz, A. J., and Gozzi, A. (2015). Functional connectivity hubs of the mouse brain. *NeuroImage*. (Cited on page 16.)

Liu, H., Zhao, R., Fang, H., Cheng, F., Fu, Y., and Liu, Y.-Y. (2017a). Entropy-based consensus clustering for patient stratification. *Bioinformatics*, 33(17):2691–2698. (Cited on page 20.)

Liu, X., Hu, A.-X., Zhao, J.-L., and Chen, F.-L. (2017b). Identification of key gene modules in human osteosarcoma by co-expression analysis weighted gene co-expression network analysis (wgcna). *Journal of cellular biochemistry*, 118(11):3953–3959. (Cited on pages 17 and 103.)

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550. (Cited on pages 3, 34, 35, and 39.)

Luo, H. M., Hu, Y., Pan, J. P., Xin, Y. R., Mi, X. N., Wang, J. H., and Gao, Q. (2018). Gene expression analysis reveals novel gene signatures between young and old adults in human prefrontal cortex. *Frontiers in Aging Neuroscience*, 10:259. (Cited on pages 17 and 103.)

Marguerat, S. and Bähler, J. (2010). Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67(4):569–579. (Cited on page 3.)

Mead, B., Berry, M., Logan, A., Scott, R. A., Leadbeater, W., and Scheven, B. A. (2015). Stem cell treatment of degenerative eye disease. *Stem cell research*, 14(3):243–257. (Cited on page 13.)

Melo, A., Monteiro, L., Lima, R. M., de Oliveira, D. M., de Cerqueira, M. D., and El-Bachá, R. S. (2011). Oxidative stress in neurodegenerative diseases: mechanisms and therapeutic perspectives. *Oxidative medicine and cellular longevity*, 2011. (Cited on page 20.)

Milenković, T., Ng, W. L., Hayes, W., and Pržulj, N. (2010). Optimal network alignment with graphlet degree vectors. *Cancer informatics*, 9:CIN–S4744. (Cited on page 16.)

Mitchell, C. M., El Jordi, O., and Yamamoto, B. K. (2019). Inflammatory mechanisms of abused drugs. *Role of Inflammation in Environmental Neurotoxicity*, 3:133. (Cited on page 169.)

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621. (Cited on page 3.)

Nygaard, S., Pedersen, P., Mikkelsen, T., Terzis, A., Tysnes, O., and Bjerkvig, R. (1994). Glioma cell invasion visualized by scanning confocal laser microscopy in an in vitro co-culture system. *Invasion & metastasis*, 15(5-6):179–188. (Cited on page 7.)

Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M. (1998). Computation with the kegg pathway database. *Biosystems*, 47(1-2):119–128. (Cited on page 18.)

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34. (Cited on page 18.)

Okaty, B. W., Sugino, K., and Nelson, S. B. (2011a). Cell type-specific transcriptomics in the brain. *Journal of Neuroscience*, 31(19):6939–6943. (Cited on pages 8 and 27.)

Okaty, B. W., Sugino, K., and Nelson, S. B. (2011b). A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS One*, 6(1):e16493. (Cited on pages 1, 8, and 27.)

Osterhoff, M., Frahnow, T., Seltmann, A., Mosig, A., Neunübel, K., Sales, S., Sampaio, J., Hornemann, S., Kruse, M., and Pfeiffer, A. (2014). Identification of gene-networks associated with specific lipid metabolites by weighted gene co-expression network analysis (wgcna). *Experimental and Clinical Endocrinology & Diabetes*, 122(03):P098. (Cited on page 17.)

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct rna sequencing. *Nature*, 461(7265):814. (Cited on page 3.)

O'Malley, M. A. and Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(1):58–68. (Cited on page 15.)

Pache, R. A., Céol, A., and Aloy, P. (2012). Netaligner—a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic acids research*, page gks446. (Cited on page 16.)

Park, C. Y., Hess, D. C., Huttenhower, C., and Troyanskaya, O. G. (2010). Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *PLoS Comput Biol*, 6(11):e1001009. (Cited on page 15.)

Pastrello, C., Pasini, E., Kotlyar, M., Otasek, D., Wong, S., Sangrar, W., Rahmati, S., and Jurisica, I. (2014). Integration, visualization and analysis of human interactome. *Biochemical and biophysical research communications*, 445(4):757–773. (Cited on page 15.)

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983. (Cited on page 86.)

Pržulj, N., Wigle, D. A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348. (Cited on pages 15, 89, and 146.)

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2

diabetes. *Nature*, 490(7418):55–60. (Cited on page 10.)

Qiu, J., Dando, O., Baxter, P. S., Hasel, P., Heron, S., Simpson, T. I., and Hardingham, G. E. (2018). Mixed-species rna-seq for elucidation of non-cell-autonomous control of gene transcription. *Nature protocols*, 13(10):2176. (Cited on pages 23, 25, 31, 82, 86, 155, 169, 170, 171, 176, 177, 180, 181, and 182.)

Qiu, J., McQueen, J., Bilican, B., Dando, O., Magnani, D., Punovuori, K., Selvaraj, B. T., Livesey, M., Haghi, G., Heron, S., et al. (2016). Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. *eLife*, 5:e20337. (Cited on pages 82 and 183.)

Rahmatallah, Y., Emmert-Streib, F., and Glazko, G. (2014). Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368. (Cited on page 17.)

Raplee, I. D., Evsikov, A. V., and de Evsikova, C. M. (2019). Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *Journal of Personalized Medicine*, 9(2):18. (Cited on page 36.)

Reverter, A., Ingham, A., Lehnert, S. A., Tan, S.-H., Wang, Y., Ratnakumar, A., and Dalrymple, B. P. (2006). Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*, 22(19):2396–2404. (Cited on page 17.)

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467. (Cited on page 9.)

Roy, N. S., Cleren, C., Singh, S. K., Yang, L., Beal, M. F., and Goldman, S. A. (2006). Functional engraftment of human es cell–derived dopaminergic neurons enriched by coculture with telomerase-immortalized midbrain astrocytes. *Nature medicine*, 12(11):1259–1268. (Cited on page 27.)

Salter, M. W. and Stevens, B. (2017). Microglia emerge as central players in brain disease. *Nature medicine*, 23(9):1018. (Cited on page 14.)

Santamaria, M., Fosso, B., Consiglio, A., De Caro, G., Grillo, G., Licciulli, F., Liuni, S., Marzano, M., Alonso-Alemany, D., Valiente, G., et al. (2012). Reference databases for taxonomic assignment in metagenomics. *Briefings in bioinformatics*, page bbs036. (Cited on page 10.)

Saritha, M., Joseph, K. P., and Mathew, A. T. (2013). Classification of mri brain images using combined wavelet entropy based spider web plots and probabilistic neural network. *Pattern Recognition Letters*, 34(16):2151–2156. (Cited on page 20.)

Schafer, D. P. and Stevens, B. (2015). Microglia function in central nervous system development and plasticity. *Cold Spring Harbor perspectives in biology*, 7(10):a020545. (Cited on page 14.)

Schöler, H. R. (2016). The potential of stem cells: An inventory. In *Humanbiotechnology as social challenge*, pages 45–72. Routledge. (Cited on page 13.)

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16. (Cited on page 2.)

Shannon, C. (1948). (1948)," a mathematical theory of communication", bell system technical journal, vol. 27, pp. 379-423 & 623-656, july & october. (Cited on pages 19, 90, and 92.)

Sharif, S. F., Hariri, R. J., Chang, V. A., Barie, P. S., Wang, R. S., and Ghajar, J. B. (1993). Human astrocyte production of tumour necrosis factor-α, interleukin-1β, and interleukin-6 following exposure to lipopolysaccharide endotoxin. *Neurological research*, 15(2):109–112. (Cited on page 171.)

Shenton, D., Smirnova, J. B., Selley, J. N., Carroll, K., Hubbard, S. J., Pavitt, G. D., Ashe, M. P., and Grant, C. M. (2006). Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. *Journal of Biological Chemistry*, 281(39):29011–29021. (Cited on page 147.)

Stevens, B. (2008). Neuron-astrocyte signaling in the development and plasticity of neural circuits. *Neurosignals*, 16(4):278–288. (Cited on page 162.)

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550. (Cited on page 19.)

Sun, K., Buchan, N., Larminie, C., and Pržulj, N. (2014). The integrated disease network. *Integrative Biology*, 6(11):1069–1079. (Cited on page 15.)

Sylvester, K. G. and Longaker, M. T. (2004). Stem cells: review and update. *Archives of Surgery*, 139(1):93–99. (Cited on page 12.)

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676. (Cited on page 13.)

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034. (Cited on page 30.)

Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., and Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147. (Cited on page 12.)

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511. (Cited on page 3.)

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43. (Cited on page 10.)

Van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284. (Cited on page 102.)

Vegeto, E., Bonincontro, C., Pollio, G., Sala, A., Viappiani, S., Nardi, F., Brusadelli, A., Viviani, B., Ciana, P., and Maggi, A. (2001). Estrogen prevents the lipopolysaccharide-induced inflammatory response in microglia. *Journal of Neu-*

roscience, 21(6):1809–1818. (Cited on page 169.)

Viader, A., Blankman, J. L., Zhong, P., Liu, X., Schlosburg, J. E., Joslyn, C. M., Liu, Q.-S., Tomarchio, A. J., Lichtman, A. H., Selley, D. E., et al. (2015). Metabolic interplay between astrocytes and neurons regulates endocannabinoid action. *Cell reports*, 12(5):798–808. (Cited on page 147.)

Volpert, G., Ben-Dor, S., Tarcic, O., Duan, J., Saada, A., Merrill, A. H., Pewzner-Jung, Y., and Futerman, A. H. (2017). Oxidative stress elicited by modifying the ceramide acyl chain length reduces the rate of clathrin-mediated endocytosis. *J Cell Sci*, 130(8):1486–1493. (Cited on pages 134 and 148.)

Wallace, Z., Rosenthal, S., Fisch, K., Ideker, T., Sasik, R., and Wren, J. (2018). On entropy and information in gene interaction networks. *Bioinformatics*. (Cited on page 20.)

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57. (Cited on page 3.)

Wang, Z., San Lucas, F. A., Qiu, P., and Liu, Y. (2014). Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection. *BMC bioinformatics*, 15(1):153. (Cited on page 183.)

Watson, M. (2006). Coxpress: differential co-expression in gene expression data. *BMC bioinformatics*, 7(1):509. (Cited on page 17.)

Wilson, A. D., Brownscombe, J. W., Krause, J., Krause, S., Gutowsky, L. F., Brooks, E. J., and Cooke, S. J. (2015). Integrating network analysis, sensor tags, and observation to understand shark ecology and behavior. *Behavioral Ecology*, page arv115. (Cited on page 16.)

Woodbury, D., Schwarz, E. J., Prockop, D. J., and Black, I. B. (2000). Adult rat and human bone marrow stromal cells differentiate into neurons. *Journal of neuroscience research*, 61(4):364–370. (Cited on page 27.)

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., Huang, G., Li, Y., Yan, Q., Wu, S., et al. (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *The ISME journal*, 5(3):414–426. (Cited on page 10.)

Xu, J., Hu, C., Chen, S., Shen, H., Jiang, Q., Huang, P., and Zhao, W. (2017). Neuregulin-1 protects mouse cerebellum against oxidative stress and neuroinflammation. *Brain research*, 1670:32–43. (Cited on page 170.)

Xu, J. and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805. (Cited on page 15.)

Xue, J., Schmidt, S. V., Sander, J., Draffehn, A., Krebs, W., Quester, I., De Nardo, D., Gohel, T. D., Emde, M., Schmidleithner, L., et al. (2014). Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity*, 40(2):274–288. (Cited on page 15.)

Yuan, L., Liu, S., Bai, X., Gao, Y., Liu, G., Wang, X., Liu, D., Li, T., Hao, A., and Wang, Z. (2016). Oxytocin inhibits lipopolysaccharide-induced inflammation in microglial cells and attenuates microglial activation in lipopolysaccharide-treated mice. *Journal of neuroinflammation*, 13(1):77. (Cited on page 169.)

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1). (Cited on page 102.)

Zhang, B., Li, H., Riggins, R. B., Zhan, M., Xuan, J., Zhang, Z., Hoffman, E. P., Clarke, R., and Wang, Y. (2009). Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*, 25(4):526–532. (Cited on page 17.)

Zhang, Z., Chen, G., Zhou, W., Song, A., Xu, T., Luo, Q., Wang, W., Gu, X.-s., and Duan, S. (2007). Regulated atp release from astrocytes through lysosome exocytosis. *Nature cell biology*, 9(8):945. (Cited on page 163.)

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., and Hovarth, S. (2010). Weighted gene coexpression network analysis: state of the art. *Journal of biopharmaceutical statistics*, 20(2):281–300. (Cited on page 102.)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`http://code.google.com/p/classicthesis/`

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

`http://postcards.miede.de/`

Final Version as of June 28, 2019 (`classicthesis` version 1.0).